

Convergence and Applications of a Gossip-based Gauss-Newton Algorithm

Xiao Li, *Student Member, IEEE*, and Anna Scaglione, *Fellow, IEEE*

Abstract—The Gauss-Newton algorithm is a popular and efficient centralized algorithm for solving non-linear least squares problems. In a large network, however, distributed observations are usually aggregated at a fusion center in order to apply the algorithm centrally, which creates inevitable communication and storage bottlenecks. In this paper, we study a distributed version of Gauss-Newton algorithm via gossiping, and show the convergence of this Gossip-based Gauss-Newton (GGN) algorithm. As an example, we show numerically that the proposed GGN algorithm is effective and robust in solving power system state estimation problems, and that the Mean Square Error (MSE) performance remains comparable to the centralized scheme and degrades gracefully even when the network exhibits random link/node failures.

Index Terms—Gauss-Newton algorithm, gossiping, distributed, convergence

I. INTRODUCTION

Numerical algorithms for solving non-linear least squares (NLLS) problems are well studied and understood [1]. The most popular solution to NLLS problems is through the so called *Newton* and *Gauss-Newton* method. Newton algorithms are second order methods that use the Hessian of the objective function to stabilize and accelerate local convergence [2], [3], while Gauss-Newton simplifies the computation of the Hessian particularly for NLLS problems by ignoring the high order derivatives [4]. The Gauss-Newton algorithm is often used in NLLS problems due to its convenience and simplicity, with applications including (but not limited to) power systems [5], localization [6], frequency estimation [7], Kalman filtering [8], medical imaging [9] and so on.

In a large network, the most common sensing architecture includes a central node that aggregates all the acquired data from the distributed agents and implement these algorithms centrally. Unfortunately, due to the scale of a networked system, centralized processing schemes do not scale well, because of the information bottleneck as well as the overhead that is required to support reliable aggregation of data. Also, these methods typically rely upon aggregation trees and are, therefore, susceptible to single-point failures, denial-of-service attacks due to network congestions and to random link or node failures. All the features that are lacking in centralized schemes can be found in gossip algorithms. Since their introduction [10], they have been extensively investigated [11], [12], as surveyed in [13]. Deterministic and randomized protocols for

gossip algorithms with synchronous or asynchronous updates have been further studied [14], [15] and applied in different areas in networked control and distributed signal processing, such as distributed Kalman filtering [16] for estimation and tracking in a network, or convex optimization problems using gossip-based sub-gradient updates [17].

Particularly relevant to gossip algorithms are recent advances made in distributed optimization via *network diffusion*, which evolved from incremental methods [18], [19] and gossip algorithms [17] onto fully decentralized and randomized algorithms. The distributed algorithms analyzed in [20]–[24] tackle convex optimization problems through either synchronous or asynchronous communications. These techniques combine a local descent step with a network diffusion step. The convergence of these diffusion algorithms typically requires *convexity* and a diminishing step-size, which results in slow convergence in general [25]. Recently, [26] proposed a diffusion optimization scheme for general non-linear convex problems with a constant step-size by assuming *local strong convexity*. Furthermore, convergence analysis of network diffusion algorithms has also been developed for adaptive signal processing on streaming observations using a constant step-size [27]–[29] for linear filtering problems, or using a diminishing step-size [24] for non-linear invertible systems. Despite the simplicity of first order methods used in diffusion algorithms, they generally suffer from slow convergence in contrast to Newton-type algorithms.

To enhance the speed of first order methods, convergence results of a gossip-based Newton method are derived in [30] for network utility maximization problems. The algorithm relies on the diagonal structure of the Hessian matrix, and this approach is later applied to power flow estimation [31]. We note, however, that the convergence of the distributed Newton method is proven under the hypothesis that the error of the computed Newton descent is bounded, which is assumed to be true but not verified. Last but not least, the method is developed specifically for *strictly convex* problems, where the variables are completely separable for each distributed agent, while many NLLS problems are oftentimes non-convex and inseparable, and thus it is unclear how these methods perform in practice if applied directly. There are also some ad-hoc applications of Gauss-Newton methods via network average consensus in sensor networks [32]–[34] or incremental methods in acoustic sources localization [35] and yet these works use gossiping to compute the algorithm update for the problems at hand, lacking the analysis of their convergence and performance.

This work was supported by the U.S. Department of Energy through the Trustworthy Cyber Infrastructure for Power Grid (TCIPG) program.

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, One Shields Avenue, Kemper Hall, Davis, California 95616-5294 (email : {eceli,ascaglione}@ucdavis.edu).

A. Contributions and Applications

The contributions of this paper are stated as follows. Firstly, we study a Gossip-based Gauss-Newton (GGN) algorithm for general NLLS problems, which exhibits much faster convergence than first order network diffusion algorithms. This algorithm only requires flexible near-neighbor communications and the communication graph can be time-varying. Secondly, we present an analysis on the local convergence and performance of the GGN algorithm, which can serve as application guidelines in a general setting. This is especially useful for problems that arise in [32]–[34], which uses a similar approach in an ad-hoc manner without discussing related numerical issues. Finally, we showcase the application of the GGN algorithm in power system state estimation [36], [37], which is a classic NLLS problem in power systems that has gained considerable interests and popularity lately to realize wide-area distributed control. We note that there has been tremendous effort spent on developing distributed state estimation schemes to alleviate the computation burden of the centralized processing [38]–[47]. Most of these algorithms distribute the computations by hierarchically aggregating or fusing the information and state estimates from distributed control areas, which assumes that there are redundant local measurements available at each area to uniquely identify the local state variables (i.e., local observability), which we do not assume in this paper.

Recently, there are methods proposed for distributed state estimation that do not require hierarchical aggregation nor local observability [48], [49]. In comparison, the proposed GGN algorithm is very different in terms of the network communications and algorithm convergence. The method in [48] is based on the diffusion algorithm motivated by [24] (similar to [20] in a non-adaptive setting), which is a first order method requiring coordination among all the agents and resulting in slow convergence. Our approach converges much faster, and the communication model we imposed, which can be either coordinated or uncoordinated, is much more flexible and robust. On the other hand, [49] follows a similar approach in [42]–[44] and uses the *alternating direction method of multipliers* (AD-MoM) to distribute the state estimation procedure by decomposing the state variables in different areas so that each agent estimates a local state, in contrast to the global state considered in this paper. Furthermore, the communications entailed by AD-MoM is constrained by the power grid topology, while the communication model considered in this paper is decoupled from the grid topology and more flexible in terms of network reconfigurations and random failures. Also, the numerical tests in [49] are based exclusively on a linear model using PMU data, while the algorithm convergence in general is not discussed. In contrast, the convergence of the proposed GGN algorithm in this paper is thoroughly analyzed. In the simulations, we show the performance of our GGN algorithm against the algorithm tailored for power system state estimation in [48] and its generalized version [24] in an adaptive setting with real-time streaming measurements.

B. Notation and Paper Organization

We denote vectors and matrices by boldface lower-case and boldface upper-case symbols and the set of real (complex) numbers by \mathbb{R} (\mathbb{C}). The magnitude of a complex number x is denoted by $|x| = \sqrt{xx^*}$, where x^* is the conjugate of the complex number x . The transpose, conjugate transpose, and inverse of a matrix \mathbf{X} are denoted by \mathbf{X}^T , \mathbf{X}^H and \mathbf{X}^{-1} , respectively. The inner products between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{N \times 1}$ is defined accordingly as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=1}^N y_n^* x_n$. The \mathbf{W} -weighted Euclidean norm of a vector \mathbf{x} is denoted by $\|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}^H \mathbf{W} \mathbf{x}}$, and the conventional Euclidean norm is written as $\|\mathbf{x}\|$. The Euclidean norm of a matrix \mathbf{A} is denoted by $\|\mathbf{A}\|$ and the *Frobenius* norm is denoted by $\|\mathbf{A}\|_F$. Given a matrix $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_N]$ where \mathbf{a}_n is a column vector, the vectorization operator is defined as $\text{vec}(\mathbf{A}) = [\mathbf{a}_1^T, \dots, \mathbf{a}_N^T]^T$.

The paper is organized as follows. In Section II, we define the NLLS problems and provide the corresponding distributed NLLS formulation in a network. Then, the proposed GGN algorithm is described in detail in Section III, with thorough convergence analysis conducted in Section IV. Furthermore, we formulate power system state estimation in Section V as a NLLS problem and solve it using the proposed GGN algorithm. Finally, the convergence and performance of the GGN algorithm for power system state estimation is demonstrated in Section VI with comparisons to other decentralized estimation schemes.

II. PROBLEM STATEMENT

A non-linear least squares (NLLS) problem is typically formulated as

$$\min \|\mathbf{g}(\mathbf{x})\|^2, \quad (1)$$

where $\mathbf{x} \in \mathbb{X}$ is the underlying N -dimensional state vector belonging to a closed convex feasible set \mathbb{X} , and $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_M(\mathbf{x})]^T$ is a vector-valued function with M outputs determined by $g_m(\cdot) : \mathbb{R}^N \rightarrow \mathbb{R}$, $m = 1, \dots, M$.

A. Centralized Gauss-Newton Algorithm

If all the agents aggregate their data and the functions to a central point, then the Gauss-Newton method is usually employed to solve NLLS problems [4] in an iterative manner as follows

$$\mathbf{x}^{k+1} = P_{\mathbb{X}} \left[\mathbf{x}^k - \alpha_k \mathbf{d}^k \right], \quad k = 1, \dots, K \quad (2)$$

with some initialization \mathbf{x}^0 , where α_k is the step-size in the k -th iteration and $P_{\mathbb{X}}[\cdot]$ is a projection onto the constrained set \mathbb{X} . The quantity \mathbf{d}^k is the Gauss-Newton descent

$$\mathbf{d}^k = [\mathbf{G}^T(\mathbf{x}^k) \mathbf{G}(\mathbf{x}^k)]^{-1} \mathbf{G}^T(\mathbf{x}^k) \mathbf{g}(\mathbf{x}^k), \quad (3)$$

where $\mathbf{G}(\mathbf{x})$ is the $M \times N$ Jacobian matrix defined as $\mathbf{G}(\mathbf{x}) = \partial \mathbf{g}(\mathbf{x}) / \partial \mathbf{x}^T$. In general, the fixed point of the algorithm update (2) is not unique, each corresponding to one of the stationary points of the cost function satisfying the first order condition

$$\mathbf{G}^T(\mathbf{x}^*) \mathbf{g}(\mathbf{x}^*) = \mathbf{0}, \quad \mathbf{x}^* \in \mathbb{X}. \quad (4)$$

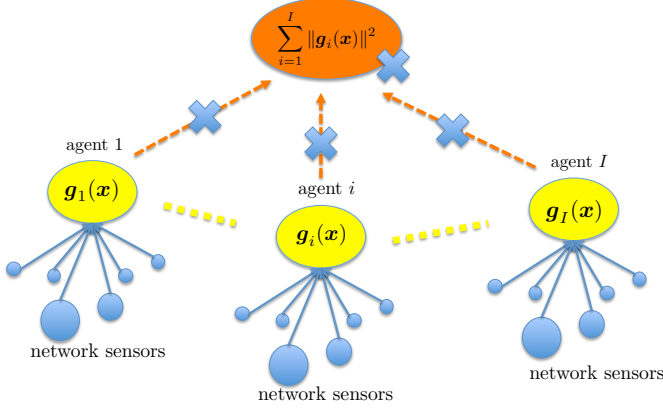


Fig. 1. Schematic of multi-agent computation structure.

Note that if α_k is chosen differently at each iteration, the algorithm is called the *damped Gauss-Newton* method while $\alpha_k = \alpha$ corresponds to the *undamped Gauss-Newton* method. If the Gauss-Newton Hessian matrix $\mathbf{G}^T(\mathbf{x}^k)\mathbf{G}(\mathbf{x}^k)$ is positive semi-definite, the resultant \mathbf{d}^k constitutes a descent direction of the objective function. It is well-known that if the step-size α_k chosen according to the Wolfe condition [1], the Gauss-Newton iteration converges to a stationary point of the cost function. Since many NLLS problems are non-convex by nature, the focus in this paper is to study the local convergence property of the algorithm to an arbitrary fixed point $\mathbf{x}^* \in \mathbb{X}$.

B. Distributed Formulation

Although the centralized Gauss-Newton algorithm is well understood, it is not immediately clear how this iteration can be implemented in a distributed manner such that similar local convergence properties can be maintained. As shown in Fig. 1, suppose there are I distributed *agents*, and the i -th *agent* only knows a subset function $\mathbf{g}_i(\mathbf{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^{M_i}$ from (1)

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \mathbf{g}_1(\mathbf{x}) \\ \vdots \\ \mathbf{g}_I(\mathbf{x}) \end{bmatrix} \quad (5)$$

with $M = \sum_{i=1}^I M_i$. In this setting, the goal is to minimize

$$\min \sum_{i=1}^I \|\mathbf{g}_i(\mathbf{x})\|^2, \quad (6)$$

where each agent has only partial knowledge of this global cost function. In this case, it is also not clear how to coordinate the step-size at different agents such that the Wolfe condition [1] would satisfy in a global sense. A variable step-size is also quite inconvenient in a decentralized setting, because of the difficulties of coordinating a change in the step-size across a network. As a result, we study the convergence behavior of the *undamped Gauss-Newton* case with a constant step-size $\alpha_k = \alpha \leq 1$ for an arbitrary agent i

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}} [\mathbf{x}_i^k - \alpha \mathbf{d}_i^k], \quad (7)$$

where the exact decentralized descent is given by

$$\mathbf{d}_i^k = [\mathbf{G}^T(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k)]^{-1} \mathbf{G}^T(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k). \quad (8)$$

It is known from (8) that at each iteration, each agent requires the computation of

$$\begin{aligned} \mathbf{G}^T(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k) &= \sum_{p=1}^I \mathbf{G}_p^T(\mathbf{x}_i^k)\mathbf{G}_p(\mathbf{x}_i^k) \\ \mathbf{G}^T(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k) &= \sum_{p=1}^I \mathbf{G}_p^T(\mathbf{x}_i^k)\mathbf{g}_p(\mathbf{x}_i^k), \end{aligned}$$

while the i -th agent has only partial information available to compute $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k)$ and $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k)$ locally. Note that these quantities are written as sums of individual network components, therefore, the distributed computations of the Gauss-Newton algorithm can emulate the centralized version by exchanging information between different agents, similar to that in [32]. Therefore, in next section, we introduce the Gossip-based Gauss-Newton (GGN) algorithm in a network for solving NLLS problems.

III. DISTRIBUTED GAUSS-NEWTON ALGORITHM VIA NETWORK GOSSIPING

The proposed GGN algorithm emulates the computation of \mathbf{d}^k in (3) in a fully distributed manner. There are two time scales in the GGN algorithm, one is the time for Gauss-Newton *update* and the other is the gossip *exchange* between every two Gauss-Newton updates. Throughout the rest of the paper, we consistently use *update* for the Gauss-Newton algorithm, denoted by the discrete time index “ k ” and *exchange* for network gossiping, denoted by another discrete time index “ ℓ ”. We assume that all the network agents have a common clock that runs synchronously and determines the time instants $t = \tau_k$ for the k -th algorithm update across the network. Between two successive updates $t \in [\tau_k, \tau_{k+1})$, the agents communicate and exchange information with each other in the form of network gossiping at time $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ for $\ell = 1, \dots, \ell_k$ after each update k . The flow chart of the algorithm is illustrated in Fig. 2, where the algorithm update and gossip exchange alternate in time.

In this section, we walk through the local update model for the k -th algorithm update at each distributed agent in Section III-A and introduce in Section III-B the gossip model for every exchange $\ell = 1, \dots, \ell_k$ that takes place between the k -th and $(k+1)$ -th updates, where the I agents advance their computations via network gossiping.

A. Local Update Model

Let \mathbf{x}_i^k be the local iterate at the i -th agent after the k -th update. For convenience, let

$$\mathbf{r}(\mathbf{x}_i^k) = \frac{1}{I} \sum_{p=1}^I \mathbf{G}_p^T(\mathbf{x}_i^k)\mathbf{g}_p(\mathbf{x}_i^k), \quad (9)$$

$$\mathbf{R}(\mathbf{x}_i^k) = \frac{1}{I} \sum_{p=1}^I \mathbf{G}_p^T(\mathbf{x}_i^k)\mathbf{G}_p(\mathbf{x}_i^k). \quad (10)$$



Fig. 2. GGN algorithm update and gossip exchange flow chart.

The “exact decentralized descent” in (8), if it were to be computed at the i -th agent for the $(k+1)$ -th update, can be equivalently obtained as

$$\mathbf{d}_i^k = \mathbf{R}(\mathbf{x}_i^k)^{-1} \mathbf{r}(\mathbf{x}_i^k), \quad (11)$$

which is impossible to obtain in a distributed setting. This is on one hand because the accessible functions $\mathbf{g}_p(\cdot)$ and Jacobians $\mathbf{G}_p(\cdot)$ to agent p does not have the local estimate \mathbf{x}_i^k at agent $i \neq p$ to evaluate their functions for such computations, and on the other hand, the i -th agent does not have the accessible functions $\mathbf{g}_p(\cdot)$ and Jacobians $\mathbf{G}_p(\cdot)$ even if they can be evaluated by other agents. In fact, the available information at the i -th agent after the k -th Gauss-Newton update is $\mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{g}_i(\mathbf{x}_i^k)$ and $\mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{G}_i(\mathbf{x}_i^k)$. Therefore alternatively, we propose to use an average surrogate for $\mathbf{r}(\mathbf{x}_i^k)$ and $\mathbf{R}(\mathbf{x}_i^k)$

$$\bar{\mathbf{h}}_k = \frac{1}{I} \sum_{i=1}^I \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{g}_i(\mathbf{x}_i^k), \quad (12)$$

$$\bar{\mathbf{H}}_k = \frac{1}{I} \sum_{i=1}^I \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{G}_i(\mathbf{x}_i^k), \quad (13)$$

which can be obtained via network gossiping. Intuitively speaking, if the distributed estimates stay close $\mathbf{x}_i^k \approx \mathbf{x}_j^k$, then $\bar{\mathbf{h}}_k \approx \mathbf{r}(\mathbf{x}_i^k)$ and $\bar{\mathbf{H}}_k \approx \mathbf{R}(\mathbf{x}_i^k)$ for all i .

Therefore, after the k -th update by each agent at τ_k , the network enters gossip exchange stage $[\tau_k, \tau_{k+1})$ to compute the surrogate $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$. Define the length- N_ϕ local information vector (i.e., $N_\phi = N(N+1)$) at the i -th agent for the ℓ -th gossip exchange at $\tau_{k,\ell}$

$$\phi_{k,i}(\ell) = \begin{bmatrix} \mathbf{h}_{k,i}(\ell) \\ \text{vec}[\mathbf{H}_{k,i}(\ell)] \end{bmatrix}, \quad (14)$$

evolving from the initial information $\phi_{k,i}(0)$ at each agent with $\mathbf{h}_{k,i}(0)$ and $\mathbf{H}_{k,i}(0)$ given below

$$\mathbf{h}_{k,i}(0) \triangleq \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{g}_i(\mathbf{x}_i^k), \quad (15)$$

$$\mathbf{H}_{k,i}(0) \triangleq \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{G}_i(\mathbf{x}_i^k). \quad (16)$$

With this definition, clearly we have $\bar{\mathbf{h}}_k = \sum_{i=1}^I \mathbf{h}_{k,i}(0)/I$ and $\bar{\mathbf{H}}_k = \sum_{i=1}^I \mathbf{H}_{k,i}(0)/I$. Then, all network agents exchange their information $\mathbf{h}_{k,i}(\ell) \rightarrow \mathbf{h}_{k,i}(\ell+1)$ and $\mathbf{H}_{k,i}(\ell) \rightarrow \mathbf{H}_{k,i}(\ell+1)$ using the gossiping protocol described later in Section III-B.

After ℓ_k exchanges, the “inexact descent” for the $(k+1)$ -th update at the i -th agent is

$$\mathbf{d}_i^k(\ell_k) = \mathbf{H}_{k,i}^{-1}(\ell_k) \mathbf{h}_{k,i}(\ell_k) \quad (17)$$

and the local estimate is updated as

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}} \left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k(\ell_k) \right]. \quad (18)$$

B. Network Gossiping Model

Before we start, we begin by describing the communication model for network gossiping that supports the information flows and exchange among the agents. We borrow the insights from [10], [20], [50] and impose some rules on the communications among all the agents over time. Let us consider a time-varying random communication graph $\mathcal{G}_{k,\ell} = (\mathcal{I}, \mathcal{M}_{k,\ell})$ during $[\tau_{k,\ell}, \tau_{k,\ell+1})$ for every k and ℓ , with the node set $\mathcal{I} = \{1, \dots, I\}$ (i.e., *agent*) and the edge set $\{i, j\} \in \mathcal{M}_{k,\ell}$. The edge set $\mathcal{M}_{k,\ell}$ is characterized by the adjacency matrix $\mathbf{A}_k(\ell) = [A_{ij}^{(k,\ell)}]_{I \times I}$, where $A_{ij}^{(k,\ell)} = 1$ if $\{i, j\} \in \mathcal{M}_{k,\ell}$ and 0 if otherwise. For each agent i in the network, there is an associated neighbor set $\mathcal{M}_{k,\ell}^{(i)} \triangleq \{j : \{i, j\} \in \mathcal{M}_{k,\ell}\}$ with which each agent exchanges information locally.

Assumption 1. The composite graph $(\mathcal{I}, \bigcup_{\ell=0}^{\infty} \mathcal{M}_{k,\ell'})$ is connected for all $\ell \geq 0$ and there exists an integer $L \geq 1$ such that for every agent pair $\{i, j\}$ in the composite graph, we have for any $\ell \geq 0$

$$\{i, j\} \in \mathcal{M}_{k,\ell} \cup \mathcal{M}_{k,\ell+1} \cup \dots \cup \mathcal{M}_{k,\ell+L-1}. \quad (19)$$

The above assumption states that the composite graph consists of all agent pairs $\{i, j\}$ that communicate directly infinitely many times and that this composite graph is connected. Furthermore, each agent communicates with another agent in the composite graph within a bounded communication interval of L , such that there exists an active link between any agent pair $\{i, j\} \in \bigcup_{\ell'=0}^{\infty} \mathcal{M}_{k,\ell'}$ at least every L consecutive time slots $[\tau_{k,\ell}, \tau_{k,\ell+L-1}] \subseteq (\tau_k, \tau_{k+1})$ for any ℓ .

With the communication model described in Assumption 1, each agent combines the information sent from its neighbors with certain weights. Define a weight matrix $\mathbf{W}_k(\ell) \triangleq [W_{ij}^k(\ell)]_{I \times I}$ for the network topology during $[\tau_{k,\ell}, \tau_{k,\ell+1})$, where the (i, j) -th entry $W_{ij}^k(\ell)$ of the matrix $\mathbf{W}_k(\ell)$ is the weight associated to the edge $\{i, j\}$, which is non-zero if and only if $\{i, j\} \in \mathcal{M}_{k,\ell}$. Therefore, the matrix $\mathbf{W}_k(\ell)$ has the same sparsity pattern as the communication network graph $\mathbf{A}_k(\ell)$, and it is determined by the network’s connectivity.

Assumption 2. For all k and ℓ , the weight matrix $\mathbf{W}_k(\ell)$ is symmetric $W_{ij}^k(\ell) = W_{ji}^k(\ell)$ and doubly stochastic. There exists a scalar η with $0 < \eta < 1$ such that for all $i, j \in \mathcal{I}$

- 1) $W_{ii}^k(\ell) \geq \eta$ for all $k > 0$ and $\ell > 0$.
- 2) $W_{ij}^k(\ell) \geq \eta$ for all $k > 0$ and $\ell > 0$ if $\{i, j\} \in \mathcal{M}_{k,\ell}$.
- 3) $W_{ij}^k(\ell) = 0$ for all $k > 0$ and $\ell > 0$ if $\{i, j\} \notin \mathcal{M}_{k,\ell}$.

The gossip exchange of each agent is local with its neighbors using this weight matrix $\mathbf{W}_k(\ell)$. For the i -th agent, the

local information is mixed with its neighbors as

$$\phi_{k,i}(\ell) = W_{ii}^k(\ell)\phi_{k,i}(\ell-1) + \sum_{j \in \mathcal{M}_{k,\ell}^{(i)}} W_{ij}^k(\ell)\phi_{k,j}(\ell-1) \quad (20)$$

for all $i = 1, \dots, I$. Then for the next exchange $\ell + 1$, the network repeats the same process. By stacking the local information vectors into an ensemble vector $\phi_k(\ell) \triangleq [\phi_{k,1}^T(\ell), \dots, \phi_{k,I}^T(\ell)]^T$, the exchange model can be written compactly as

$$\phi_k(\ell) = [\mathbf{W}_k(\ell) \otimes \mathbf{I}_{N_\phi}] \phi_k(\ell-1), \quad 1 \leq \ell \leq \ell_k, \quad (21)$$

where ℓ_k is the maximum number of message exchanges during $[\tau_k, \tau_{k+1})$.

The gossip exchange model specified in (20) under Assumption 1 and 2 is a general model that includes time-varying network formations, where all agents form random communication links with their neighbors and advance their computations of the average of all local information vectors $\bar{\phi}_k = \sum_{i=1}^I \phi_{k,i}(0)/I$. With the prescribed communication model, we highlight the following two special cases which are often analyzed in consensus and gossiping literature [10], [13], [15], [17].

1) **Coordinated Static Exchange (CSE)** [13], [17]: In the CSE protocol, each agent combines the information from possible multiple neighbors, determined by the communication network \mathbf{A} , with a static weight matrix \mathbf{W} for all updates and exchanges at $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ for $\ell = 1, \dots, \ell_k$. In particular, if the network is fully connected such that $\mathbf{A} = \mathbf{I} - \mathbf{1}\mathbf{1}^T$, the communication interval is simply $L = 1$ in which each agent talks to everybody in every exchange. There are multiple ways to choose the weight matrix in the CSE protocol, where one of the most popular choice is constructed according to the Laplacian $\mathbf{L} = \text{diag}(\mathbf{A}\mathbf{1}_I) - \mathbf{A}$ as $\mathbf{W} = \mathbf{I}_I - w\mathbf{L}$ with $w = \beta/\max(\mathbf{A}\mathbf{1}_I)$ for some $0 < \beta < 1$.

2) **Uncoordinated Random Exchange (URE)** [15]: For each exchange in the URE protocol during $[\tau_k, \tau_{k+1})$, a random agent i wakes up and chooses at random a neighbor agent $j \in \mathcal{M}_{k,\ell}^{(i)}$ to communicate. We define the matrix $\Gamma \triangleq [\gamma_{i,j}]_{I \times I}$ whose (i,j) -th element $\gamma_{i,j}$ represents the probability of node i choosing agent j once agent i wakes up. The gossip exchanges are pairwise and local [15]. Suppose agent $I_{k,\ell}$ wakes up at $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ and $J_{k,\ell}$ is the node picked by node $I_{k,\ell}$ with probability $\gamma_{I_{k,\ell}, J_{k,\ell}}$. Then given some mixing parameter $0 < \beta < 1$, the weight matrix at this time is

$$\mathbf{W}_k(\ell) = \mathbf{I} - \beta (\mathbf{e}_{I_{k,\ell}} + \mathbf{e}_{J_{k,\ell}}) (\mathbf{e}_{I_{k,\ell}} + \mathbf{e}_{J_{k,\ell}})^T. \quad (22)$$

We acknowledge that typically, the URE protocol is completely random and asynchronous, which does not necessarily satisfy Assumption 1 due to link failures and random link formation between the communicating agents. In the analysis, we assume that Assumption 1 holds, while in the simulations we show the robustness of the proposed algorithm to link failures when Assumption 1 may not hold.

To facilitate further study on the network gossiping error under the general model, we invoke the following lemma and use it in later analysis.

Lemma 1. [20, Proposition 1] *Let Assumption 1 and 2 hold. Then the entries of the product $\left[\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell')\right]_{ij}$ for every k converge with a geometric rate uniformly for all $i, j \in \mathcal{I}$ as*

$$\left| \left[\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') \right]_{ij} - \frac{1}{I} \right| \leq 2 \frac{1 + \eta^{-L_0}}{1 - \eta^{L_0}} (1 - \eta^{L_0})^{\ell/L_0}, \quad (23)$$

with $L_0 \geq L$ being the least number of exchange for every agent to communicate exhaustively with all other agents in the composite graph, and the limit exists when $\ell \rightarrow \infty$

$$\lim_{\ell \rightarrow \infty} \prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') = \frac{1}{I} \mathbf{1}\mathbf{1}^T. \quad (24)$$

Given Lemma 1, we have

$$\lim_{\ell \rightarrow \infty} \phi_{k,i}(\ell) = \frac{1}{I} \sum_{i=1}^I \phi_{k,i}(0), \quad (25)$$

which accordingly results in an asymptotic local descent

$$\lim_{\ell \rightarrow \infty} \mathbf{d}_i^k(\infty) = \bar{\mathbf{H}}_k^{-1} \bar{\mathbf{h}}_k. \quad (26)$$

Note that the error made in computing the local descent (17) to emulate the exact descent in (11) stems from two sources, including the gossiping error that persists with a finite ℓ_k and the mismatch error by using the average surrogates $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$ instead of the exact global information. In the following section, we analyze the effect of this error in the convergence of the decentralized algorithm.

IV. CONVERGENCE ANALYSIS

The GGN algorithm is summarized in **Algorithm 1**. In this section, we analyze the convergence of the local GGN algorithm by examining the recursion in (18). We make the following generic assumptions on the NLLS problem.

Assumption 3. *We assume the following properties*

- 1) *The vector function is bounded $\|\mathbf{g}(\mathbf{x})\| \leq C_g$ for $\mathbf{x} \in \mathbb{X}$.*
- 2) *Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the minimum and maximum eigenvalues and let*

$$\sigma_{\min} = \min_{\mathbf{x} \in \mathbb{X}} \sqrt{\lambda_{\min}(\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x}))},$$

$$\sigma_{\max} = \max_{\mathbf{x} \in \mathbb{X}} \sqrt{\lambda_{\max}(\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x}))}.$$

Assume that the Jacobian $\mathbf{G}(\mathbf{x})$ is full-column rank for all $\mathbf{x} \in \mathbb{X}$ with $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$.

- 3) *The Jacobian $\mathbf{G}(\mathbf{x})$ satisfies the Lipschitz condition*

$$\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\| \leq \omega \|\mathbf{x} - \mathbf{x}'\|, \quad \mathbf{x}, \mathbf{x}' \in \mathbb{X},$$

where ω is the Lipschitz constant.

A. Perturbed Error Recursion Analysis

At the $(k+1)$ -th update, the error between the local estimate \mathbf{x}_i^{k+1} and a fixed point in (4) satisfies the following recursion.

Lemma 2. *Let \mathbb{X} be a closed convex set and Assumption 3 hold. The error $\|\mathbf{x}_i^{k+1} - \mathbf{x}^*\|$ between the local iterate \mathbf{x}_i^k*

Algorithm 1 Gossip-based Gauss-Newton (GGN) Algorithm

- 1: **given** initial variables \mathbf{x}_i^0 at all agents $i \in \mathcal{I}$.
- 2: **set** $k = 0$.
- 3: **repeat**
- 4: **set** $k = k + 1$.
- 5: **initialization:** For $i \in \mathcal{I}$, each agent i evaluates

$$\mathbf{h}_{k,i}(0) = \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{g}_i(\mathbf{x}_i^k) \quad (27)$$

$$\mathbf{H}_{k,i}(0) = \mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{G}_i(\mathbf{x}_i^k), \quad (28)$$

and constructs $\phi_{k,i}(0)$ as (14);

- 6: **network gossiping:** Each agent i exchanges information with neighbors via gossiping

$$\phi_k(\ell) = [\mathbf{W}_k(\ell) \otimes \mathbf{I}_{N_\phi}] \phi_k(\ell - 1), \quad 1 \leq \ell \leq \ell_k.$$

- 7: **local update:** For $i \in \mathcal{I}$, each agent i updates

$$\mathbf{d}_i^k(\ell_k) = \mathbf{H}_{k,i}^{-1}(\ell_k) \mathbf{h}_{k,i}(\ell_k) \quad (29)$$

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}} \left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k(\ell_k) \right] \quad (30)$$

- 8: **until** $\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\| \leq \epsilon$ or $k = K$.
- 9: **set** the local estimate as $\hat{\mathbf{x}}_i = \mathbf{x}_i^k$.

generated by the algorithm update (18) and an arbitrary fixed point \mathbf{x}^* in (4) satisfies the following recursion

$$\|\mathbf{x}_i^{k+1} - \mathbf{x}^*\| \leq T_1 \|\mathbf{x}_i^k - \mathbf{x}^*\|^2 + T_2 \|\mathbf{x}_i^k - \mathbf{x}^*\| \quad (31)$$

$$+ \alpha \|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|, \quad (32)$$

where $T_1 \triangleq \omega/2\sigma_{\min}$, $T_2 \triangleq (1 - \alpha)\sigma_{\max}/\sigma_{\min} + \alpha\sqrt{2}\omega\epsilon_*/\sigma_{\min}^2$ and $\epsilon_* = \|\mathbf{g}(\mathbf{x}^*)\|$.

Proof: See Appendix A. ■

The error recursion is a perturbed version of the centralized recursion with the discrepancy error $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|$ between the distributed and centralized update. Next, we propose the main result on the convergence of this perturbed recursion.

Theorem 1. (Sufficient Condition for Convergence with Bounded Perturbation) Let Assumption 3 hold. Let $\mathbf{x} \in \mathbb{X}$ be a closed convex set and

$$\rho_{\min} = \frac{(1 - T_2) - \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \quad (33)$$

$$\rho_{\max} = \frac{(1 - T_2) + \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \quad (34)$$

with $T_1, T_2 > 0$ given in (31) for some α and $0 < \kappa \leq (1 - T_2)^2/4\alpha T_1$. If $\sqrt{2}\omega\epsilon_* \ll \sigma_{\min}^2$ and $(1 - \sigma_{\min}/\sigma_{\max}) < \alpha \leq 1$ such that $0 < T_2 < 1$, and the discrepancy error can be bounded as $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| \leq \kappa$ for all $k > 0$ and $i \in \mathcal{I}$, then given any \mathbf{x}_i^0 that falls within the ρ_{\max} -neighborhood of a certain fixed point $\mathbf{x}^* \in \mathbb{X}$

$$\|\mathbf{x}_i^0 - \mathbf{x}^*\| < \rho_{\max}, \quad (35)$$

the asymptotic error of the local iterate \mathbf{x}_i^k at each agent with

respect to \mathbf{x}^* can be bounded as

$$\limsup_{k \rightarrow \infty} \|\mathbf{x}_i^{k+1} - \mathbf{x}^*\| \leq \rho_{\min}. \quad (36)$$

Proof: If the discrepancy error is upper bounded by a constant $\kappa \geq 0$ such that

$$\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| \leq \kappa, \quad (37)$$

then from Lemma 2, the recursion can be upper bounded by

$$\|\mathbf{x}_i^{k+1} - \mathbf{x}^*\| \leq T_1 \|\mathbf{x}_i^k - \mathbf{x}^*\|^2 + T_2 \|\mathbf{x}_i^k - \mathbf{x}^*\| + \alpha\kappa. \quad (38)$$

Let $\zeta_{i,k} = \|\mathbf{x}_i^k - \mathbf{x}^*\|$, then the error recursion dynamics can be expressed as a dynamical system as

$$\zeta_{i,k+1} \leq T_1 \zeta_{i,k}^2 + T_2 \zeta_{i,k} + \alpha\kappa, \quad \zeta_{i,k} > 0. \quad (39)$$

Since $\zeta_{i,k}$ is non-negative, this error dynamic can be upper bounded by the dynamical system of $\rho_{k+1} = \psi(\rho_k)$ with

$$\psi(\rho_k) = T_1 \rho_k^2 + T_2 \rho_k + \alpha\kappa, \quad \rho_k > 0, \quad (40)$$

whose equilibrium points are evaluated as

$$\rho = T_1 \rho^2 + T_2 \rho + \alpha\kappa. \quad (41)$$

When κ satisfies $\kappa \leq (1 - T_2)^2/4\alpha T_1$, the equilibrium solutions to (41) exist and are obtained as in (33). According to [51], an equilibrium point is a stable sink if $|\dot{\psi}(\cdot)| < 1$ and unstable otherwise. Thus when $0 < T_2 < 1$, the equilibrium point ρ_{\min} is a sink because

$$|\dot{\psi}(\rho_{\min})| = |2T_1 \rho_{\min} + T_2| \quad (42)$$

$$= \left| 1 - \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa} \right| < 1, \quad (43)$$

where $\dot{\psi}(\rho) \triangleq d\psi(\rho)/d\rho$ is the first order derivative of the dynamics, while ρ_{\max} is unstable since $|\dot{\psi}(\rho_{\max})| > 1$ for any T_2 . To guarantee $0 < T_2 < 1$, it requires

$$\left(1 - \frac{\sigma_{\min}}{\sigma_{\max}} \right) \left(1 - \frac{\sqrt{2}\omega\epsilon_*}{\sigma_{\min}\sigma_{\max}} \right)^{-1} < \alpha \leq 1. \quad (44)$$

To guarantee that the term on the left is strictly less than 1, it is required that $\sqrt{2}\omega\epsilon_*/\sigma_{\min}^2 < 1$. Therefore given $\sqrt{2}\omega\epsilon_* \ll \sigma_{\min}^2$, it is sufficient to have $(1 - \sigma_{\min}/\sigma_{\max}) < \alpha \leq 1$ such that all the requirements are met.

Thus, if the initial error $\zeta_{i,0} > \rho_{\max}$, the error is infinitely accumulated and become unstable $\lim_{k \rightarrow \infty} \zeta_{i,k} = \infty$. On the other hand, as long as the errors are constantly bounded by

$$0 < \zeta_{i,k} < \rho_{\max}$$

for all i 's and k 's, the algorithm reaches the equilibrium error floor ρ_{\min} . Thus, as long as the initialization error $\zeta_{i,0}$ satisfies $0 < \zeta_{i,0} < \rho_{\max}$ for $i = 1, \dots, I$, the algorithm progresses with contracting error until reaching the error floor ρ_{\min} due to the constant bounded perturbation κ . Finally, it is shown that as long as the initial condition \mathbf{x}_i^0 satisfies $\|\mathbf{x}_i^0 - \mathbf{x}^*\| < \rho_{\max}$ with respect to a certain fixed point \mathbf{x}^* , the error norm is upper

bounded by the steady state error ρ_{\min}

$$\limsup_{k \rightarrow \infty} \|\mathbf{x}_i^k - \mathbf{x}^*\| \leq \rho_{\min}.$$

■

An immediate result of Theorem 1 is that if the gossip exchanges ℓ_k 's are moderately large, then $\kappa \rightarrow 0$ and the GGN iterate \mathbf{x}_i^k asymptotically approaches the centralized version \mathbf{x}^* . An intuition that can be drawn from the sufficient condition is that the smaller the Lipschitz constant ω , the larger radius off the fixed point \mathbf{x}^* can be allowed for convergence. In other words, the less non-linear the cost function the better the convergence. Furthermore, if the problem is very consistent, meaning ϵ_* is very small at stationary points (which is often the case in data fitting problems in sensor networks), then $\rho_{\max} \approx 2\sigma_{\min}/\omega - \kappa$ and the steady state error is approximately $\rho_{\min} \approx \kappa$, which scales with the errors resulted from decentralized updates via gossiping. In particular, the convergence is quadratic if $\epsilon_* = 0$ and $\kappa \rightarrow 0$, which means that the GGN algorithm achieves the same convergence rate as that in Newton's method without computing second order derivatives. Now that we have proved convergence of the GGN algorithm under bounded perturbations, we proceed to prove that the bounded perturbation assumption holds.

B. Perturbation Analysis of κ

Preceding analysis has proven that if the perturbation error due to the distributed updates $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|$ is bounded, the condition proposed in Theorem 1 is sufficient to guarantee convergence of the GGN algorithm. In the following, we proceed to analyze this perturbation and further show that the bounded condition holds when the distributed update is achieved via network gossiping.

1) **Network gossiping error:** Define at the ℓ -th exchange

$$\begin{aligned} \mathbf{h}_k(\ell) &\triangleq [\mathbf{h}_{k,1}^T(\ell), \dots, \mathbf{h}_{k,I}^T(\ell)]^T, \\ \mathbf{H}_k(\ell) &\triangleq [\mathbf{H}_{k,1}^T(\ell), \dots, \mathbf{H}_{k,I}^T(\ell)]^T \end{aligned}$$

and their deviations from the exact averages $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$ as

$$\begin{aligned} \mathbf{e}_k(\ell) &= [\mathbf{e}_{k,1}^T(\ell), \dots, \mathbf{e}_{k,I}^T(\ell)]^T, \\ \mathbf{E}_k(\ell) &= [\mathbf{E}_{k,1}^T(\ell), \dots, \mathbf{E}_{k,I}^T(\ell)]^T, \end{aligned}$$

where $\mathbf{e}_{k,i}(\ell) = \mathbf{h}_{k,i}(\ell) - \bar{\mathbf{h}}_k$ and $\mathbf{E}_{k,i}(\ell) = \mathbf{H}_{k,i}(\ell) - \bar{\mathbf{H}}_k$. The gossip errors $\mathbf{e}_k(\ell_k)$ and $\mathbf{E}_k(\ell_k)$ are closely related to the properties of the weight matrices $\mathbf{W}_k(\ell)$ given in Lemma 1, as stated below.

Lemma 3. *Let Assumption 1 and 2 hold. The gossip error $\mathbf{e}_k(\ell_k)$ and $\mathbf{E}_k(\ell_k)$ after the k -th update at the ℓ_k -th exchange can be bounded as*

$$\|\mathbf{e}_k(\ell_k)\| < C\lambda_\eta^{\ell_k}, \quad \|\mathbf{E}_k(\ell_k)\|_F < C\lambda_\eta^{\ell_k},$$

where $C = 2I^2 \sqrt{I\sigma_{\max}^2(C_g^2 + N\sigma_{\max}^2)}(1 + \eta^{-L_0})/(1 - \eta^{L_0})$ and $\lambda_\eta = (1 - \eta^{L_0})^{1/L_0}$ with $0 < \lambda_\eta < 1$.

Proof: See Appendix B. ■

2) **Perturbation:** Define the errors between the surrogate $\bar{\mathbf{h}}_k$, $\bar{\mathbf{H}}_k$ and the exact information $\mathbf{r}(\mathbf{x}_i^k)$ and $\mathbf{R}(\mathbf{x}_i^k)$ as

$$\delta_{k,i} = \bar{\mathbf{h}}_k - \mathbf{r}(\mathbf{x}_i^k) = \frac{1}{I} \sum_{p=1}^I [\mathbf{h}_{k,i}(\ell) - \mathbf{h}_{k,p}(\ell)] \quad (45)$$

$$\Delta_{k,i} = \bar{\mathbf{H}}_k - \mathbf{R}(\mathbf{x}_i^k) = \frac{1}{I} \sum_{p=1}^I [\mathbf{H}_{k,i}(\ell) - \mathbf{H}_{k,p}(\ell)].$$

Since the individual Jacobian $\mathbf{G}_i(\mathbf{x})$ is a sub-matrix of $\mathbf{G}(\mathbf{x})$, thus from the Lipschitz condition (3) in Assumption 3 and [52, Corollary 3.1.3], it also satisfies

$$\begin{aligned} \|\mathbf{G}_i(\mathbf{x}) - \mathbf{G}_i(\mathbf{x}')\| &\leq \|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\| \\ &\leq \omega \|\mathbf{x} - \mathbf{x}'\|, \quad \mathbf{x}, \mathbf{x}' \in \mathbb{X}. \end{aligned}$$

From conditions (1) and (2) in Assumption 3 and [53, Theorem 12.4], $\mathbf{G}_i^T(\mathbf{x})\mathbf{g}_i(\mathbf{x})$ and $\mathbf{G}_i^T(\mathbf{x})\mathbf{G}_i(\mathbf{x})$ also satisfy the Lipschitz condition with constants ν_δ and ν_Δ for arbitrary $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$

$$\begin{aligned} \|\mathbf{G}_i^T(\mathbf{x})\mathbf{g}_i(\mathbf{x}) - \mathbf{G}_i^T(\mathbf{x}')\mathbf{g}_i(\mathbf{x}')\| &\leq \nu_\delta \|\mathbf{x} - \mathbf{x}'\| \\ \|\mathbf{G}_i^T(\mathbf{x})\mathbf{G}_i(\mathbf{x}) - \mathbf{G}_i^T(\mathbf{x}')\mathbf{G}_i(\mathbf{x}')\| &\leq \nu_\Delta \|\mathbf{x} - \mathbf{x}'\|, \end{aligned}$$

where $\nu_\delta \geq \omega(C_g + \sigma_{\max})$ and $\nu_\Delta \geq 2\sigma_{\max}\omega$. By definition (45) and the Lipschitz conditions, we have

$$\|\delta_{k,i}\| \leq \frac{\nu_\delta}{I} \sum_{p=1}^I \|\mathbf{x}_i^k - \mathbf{x}_p^k\|, \quad (46)$$

$$\|\Delta_{k,i}\| \leq \frac{\nu_\Delta}{I} \sum_{p=1}^I \|\mathbf{x}_i^k - \mathbf{x}_p^k\|. \quad (47)$$

Thus, we can express the local information for each agent during gossip in relation to the exact information

$$\mathbf{h}_{k,i}(\ell) = \mathbf{r}(\mathbf{x}_i^k) + \delta_{k,i} + \mathbf{e}_{k,i}(\ell), \quad (48)$$

$$\mathbf{H}_{k,i}(\ell) = \mathbf{R}(\mathbf{x}_i^k) + \Delta_{k,i} + \mathbf{E}_{k,i}(\ell). \quad (49)$$

Clearly, this discrepancy depends on the gossip errors $\mathbf{e}_{k,i}(\ell_k)$ and $\mathbf{E}_{k,i}(\ell_k)$, and the inconsistency $\Delta_{k,i}$ and $\delta_{k,i}$ due to the disagreement $\|\mathbf{x}_i^k - \mathbf{x}_j^k\|$ for each pair of i -th and j -th agents. Gossip errors have been specified in Lemma 3, thus in the following we bound the disagreement $\|\mathbf{x}_i^k - \mathbf{x}_j^k\|$.

Assumption 4. *Denote the minimum number of gossip exchanges as ℓ_* = $\min_k \{\ell_k\}$. We assume that $\{\ell_k\}_{k=0}^\infty$ are chosen to satisfy¹*

$$D = \lim_{k \rightarrow \infty} \sum_{\ell=\ell_0}^{\ell_k} \lambda_\eta^{(\ell-\ell_*)} < \infty.$$

and for any given $0 < \kappa \ll 1$, the minimum exchange ℓ_* satisfies

$$\ell_* \geq \log \left(\frac{\kappa}{CC_\sigma(1 + DC_\sigma)} \right) / \log \lambda_\eta,$$

¹A simple choice is $\ell_0 = \ell_*$ and $\ell_k = \ell_{k-1} + 1$, then $D = 1/(1 - \lambda_\eta)$.

where C is a constant defined in Lemma 3 and

$$C_\sigma = \nu \left(\frac{\sigma_{\max} C_g I}{\sigma_{\min}^4} + \frac{I}{\sigma_{\min}^2} \right) \quad (50)$$

with $\nu = \max\{2, \nu_\delta, \nu_\Delta\}$.

Lemma 4. Given Lemma 1 under Assumption 1, 2, 3 and 4, if the initializer at each agent satisfies $\mathbf{x}_i^0 = \mathbf{x}^0$ for all i , the deviation $\|\mathbf{x}_i^k - \mathbf{x}_j^k\|$ for any i and j satisfies

$$\|\mathbf{x}_i^k - \mathbf{x}_j^k\| \leq CC_\sigma \lambda_\eta^{\ell_*} D_{k-1},$$

where $D_k = \sum_{\ell=\ell_0}^{\ell_k} \lambda_\eta^{(\ell-\ell_*)}$.

Proof: See Appendix C. ■

Proposition 1. Given Lemma 1, 3 and 4 under Assumption 1, 2, 3 and 4, the discrepancy between the inexact decentralized descent and the exact decentralized descent is bounded as

$$\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| < \kappa \quad (51)$$

for all i and k with an arbitrary $0 < \kappa \ll 1$.

Proof: See Appendix D. ■

Remark 1: We acknowledge that the condition in Theorem 1 is proven using very pessimistic bounds, such as the eigenvalues bounds in Assumption 3, and therefore it is sufficient but not necessary. In fact, as shown in Section VI, even when the sufficient conditions are not satisfied, the algorithm behaves well even with link failure present.

The GGN algorithm and its variants can be found in various papers, where existing works have directly used this algorithm in sensor networks [32]–[34] for localization, however without the performance and convergence analysis. The analysis conducted in this paper can serve as a general guideline to predict the numerical stability in different situations. In the following, we demonstrate its application in power systems, which has gained considerable interests and popularity in the past few years to realize wide-area distributed control.

V. APPLICATION : POWER SYSTEM STATE ESTIMATION

A power network is characterized by vertices (called *buses*) representing simple interconnections, generators or loads, denoted by the set $\mathcal{N} \triangleq \{1, \dots, N\}$. Transmission lines connect these buses are the edges of the power grid topology, denoted $\mathcal{E} \triangleq (n, l)$ with cardinality $|\mathcal{E}| = L$ that corresponds to the transmission line between bus n and l , $n, l \in \mathcal{N}$. Each transmission line is characterized by the admittance matrix $\mathbf{Y} = [-Y_{nl}]_{N \times N}$, where $Y_{nl} = G_{nl} + jB_{nl}$, $(n, l) \in \mathcal{E}$ is the line admittance, and each bus has a shunt admittance $\bar{Y}_{nl} = \bar{G}_{nl} + j\bar{B}_{nl}$ associated with the Π -model² of the transmission line $(n, l) \in \mathcal{E}$. Note that $Y_{nn} = -\sum_{l \neq n} (\bar{Y}_{nl} + Y_{nl})$ is defined as the self-admittance. The state of the power system corresponds to the voltage phasors at all buses, described by voltage phase and magnitude $\mathbf{x} = [\boldsymbol{\Theta}^T, \mathbf{V}^T]^T$, where $\boldsymbol{\Theta} \triangleq [\theta_1, \dots, \theta_N]^T$ is the phase vector with θ_1 being the slack bus phase, and $\mathbf{V} \triangleq [V_1, \dots, V_N]^T$ contains the magnitude.

² The Π -model is a circuit equivalent of a transmission line by abstracting two electric buses as a two-port network in the shape of a Π connection [54].

A. Power Measurement Models

Nowadays, power measurements include *phasor measurements* (V_n, θ_n) and the *active/reactive power injection* (P_n, Q_n) for buses $n \in \mathcal{N}$, and the *active/reactive current* (I_{nl}, J_{nl}) and the *active/reactive power flows* (P_{nl}, Q_{nl}) on transmission lines $(n, l) \in \mathcal{E}$. The ensemble of these quantities can be stacked into the length- $2N$ phasor measurement vector $\mathbf{f}_{\mathcal{X}}(\mathbf{x}) = \mathbf{x}$ and power injection vector $\mathbf{f}_{\mathcal{N}}(\mathbf{x})$, as well as the length- $4L$ current vector $\mathbf{f}_{\mathcal{I}}(\mathbf{x})$ and line flow vector $\mathbf{f}_{\mathcal{E}}(\mathbf{x})$ respectively

$$\mathbf{f}_{\mathcal{I}}(\mathbf{x}) = [\dots, I_{nl}(\mathbf{x}), \dots, \dots, J_{nl}(\mathbf{x}), \dots]^T \quad (52)$$

$$\mathbf{f}_{\mathcal{N}}(\mathbf{x}) = [P_1(\mathbf{x}), \dots, P_N(\mathbf{x}), Q_1(\mathbf{x}), \dots, Q_N(\mathbf{x})]^T \quad (53)$$

$$\mathbf{f}_{\mathcal{E}}(\mathbf{x}) = [\dots, P_{nl}(\mathbf{x}), \dots, \dots, Q_{nl}(\mathbf{x}), \dots]^T \quad (54)$$

and expressed in relation to the state \mathbf{x} as in [54]

$$P_n(\mathbf{x}) = V_n \sum_{l \neq n}^N V_l (G_{nl} \cos \theta_{nl} + B_{nl} \sin \theta_{nl})$$

$$Q_n(\mathbf{x}) = V_n \sum_{l \neq n}^N V_l (G_{nl} \sin \theta_{nl} - B_{nl} \cos \theta_{nl})$$

$$P_{nl}(\mathbf{x}) = V_n^2 G_{nl} - V_n V_l (G_{nl} \cos \theta_{nl} + B_{nl} \sin \theta_{nl})$$

$$Q_{nl}(\mathbf{x}) = -V_n^2 B_{nl} - V_n V_l (G_{nl} \sin \theta_{nl} - B_{nl} \cos \theta_{nl}),$$

where $\theta_{nl} = \theta_n - \theta_l$. The current expressions are omitted because it is directly the Ohm's law. Note that these quantities can also be expressed in terms of the real and imaginary parts of the voltage, but here we stick to the conventional polar representation of the magnitude V_n and phase θ_n .

The above power quantities are recorded by field devices in the power grid, and gathered centrally by aggregation to control centers. By stacking all these quantities together as a vector function of the state \mathbf{x} and all the measurements into a vector \mathbf{z} , the ensemble of all power measurements is

$$\mathbf{z} = \mathbf{f}(\bar{\mathbf{x}}) + \boldsymbol{\varepsilon}, \quad (55)$$

where $\bar{\mathbf{x}}$ represents the true state and

$$\mathbf{z} \triangleq \begin{bmatrix} \mathbf{z}_{\mathcal{X}} \\ \mathbf{z}_{\mathcal{I}} \\ \mathbf{z}_{\mathcal{N}} \\ \mathbf{z}_{\mathcal{E}} \end{bmatrix}, \quad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \mathbf{f}_{\mathcal{X}}(\mathbf{x}) \\ \mathbf{f}_{\mathcal{I}}(\mathbf{x}) \\ \mathbf{f}_{\mathcal{N}}(\mathbf{x}) \\ \mathbf{f}_{\mathcal{E}}(\mathbf{x}) \end{bmatrix}, \quad \boldsymbol{\varepsilon} \triangleq \begin{bmatrix} \boldsymbol{\varepsilon}_{\mathcal{X}} \\ \boldsymbol{\varepsilon}_{\mathcal{I}} \\ \boldsymbol{\varepsilon}_{\mathcal{N}} \\ \boldsymbol{\varepsilon}_{\mathcal{E}} \end{bmatrix}. \quad (56)$$

B. Formulation and Solution

In practice, a reasonable abstraction of the data acquisition architecture in power systems is as an interconnected multi-site infrastructure, with I sites in which the i -th site covers a subset of buses $n \in \mathcal{N}_i$ satisfying $\mathcal{N}_j \cap \mathcal{N}_i = \emptyset$ and $\mathcal{N}_i, \mathcal{N}_j \subset \mathcal{N}$. The i -th site temporally aligns and aggregates a snapshot of M_i local measurements of $\{z_{i,m}\}_{m=1}^{M_i}$ within the site or on the lines that connect its own site with others, as shown in Fig. 3. Also, the observations gathered are not exhaustive. The i -th site's measurements are selected from the ensemble in (55) as

$$\mathbf{z}_i = \mathbf{T}_i \mathbf{z} = \mathbf{f}_i(\mathbf{x}) + \boldsymbol{\varepsilon}_i, \quad (57)$$

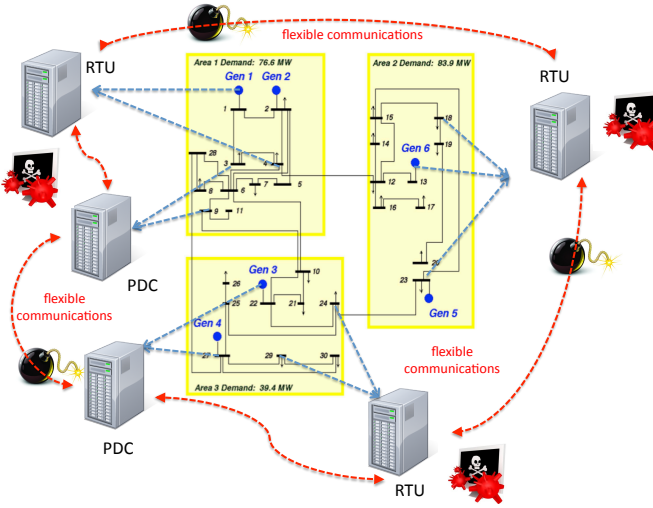


Fig. 3. Multi-site structure in IEEE-30 test case

where $\mathbf{z}_i \triangleq [\mathbf{z}_{i,\mathcal{X}}^T, \mathbf{z}_{i,\mathcal{I}}^T, \mathbf{z}_{i,\mathcal{N}}^T, \mathbf{z}_{i,\mathcal{E}}^T]^T$, $\mathbf{f}_i(\mathbf{x}) \triangleq \mathbf{T}_i \mathbf{f}(\mathbf{x})$, and $\mathbf{T}_i \triangleq \text{diag}[\mathbf{T}_{i,\mathcal{X}}, \mathbf{T}_{i,\mathcal{I}}, \mathbf{T}_{i,\mathcal{N}}, \mathbf{T}_{i,\mathcal{E}}]$ is a block diagonal binary matrix selecting the corresponding measurements at the i -th site. Specifically, $\mathbf{T}_{i,\mathcal{X}} \in \{0, 1\}^{M_{i,\mathcal{X}} \times 2N}$, $\mathbf{T}_{i,\mathcal{I}} \in \{0, 1\}^{M_{i,\mathcal{I}} \times 4L}$, $\mathbf{T}_{i,\mathcal{N}} \in \{0, 1\}^{M_{i,\mathcal{N}} \times 2N}$ and $\mathbf{T}_{i,\mathcal{E}} \in \{0, 1\}^{M_{i,\mathcal{E}} \times 4L}$ are selection matrices with each row having only one “1” entry located at the index of the corresponding element in $\mathbf{f}(\mathbf{x})$ measured by field devices. Note that the number of measurements recorded by each agent is $M_i = M_{i,\mathcal{X}} + M_{i,\mathcal{I}} + M_{i,\mathcal{N}} + M_{i,\mathcal{E}} = \text{Tr}(\mathbf{T}_i^T \mathbf{T}_i)$.

The universally accepted problem formulation for static state estimation is to solve a weighted NLLS problem that fits the estimated state to the power measurements [36], [37]. In the following we set the weights to be equal for simplicity. Assuming $\mathbb{E}\{\varepsilon\varepsilon^T\} = \sigma^2 \mathbf{I}$, the state is estimated as

$$\hat{\mathbf{x}} = \min_{\mathbf{x} \in \mathbb{X}} \sum_{i=1}^I \|\mathbf{z}_i - \mathbf{f}_i(\mathbf{x})\|^2 \quad (58)$$

where $\mathbb{X} \triangleq \{\theta_n \in [-\theta_{\max}, \theta_{\max}], V_n \in (0, V_{\max}], n \in \mathcal{N}\}$ with θ_{\max} and V_{\max} being the phase angle and voltage limit. Clearly, by letting

$$\mathbf{g}_i(\mathbf{x}) \triangleq \mathbf{z}_i - \mathbf{f}_i(\mathbf{x}), \quad \mathbf{G}_i(\mathbf{x}) \triangleq -\frac{\partial \mathbf{f}_i(\mathbf{x})}{\partial \mathbf{x}^T} \quad (59)$$

the problem can be solved using the proposed GGN algorithm.

Remark 2: In many practical scenarios [24], [27]–[29], many similar NLLS problems in a network take the form

$$\hat{\mathbf{x}} = \min_{\mathbf{x} \in \mathbb{X}} \sum_{i=1}^I \|\mathbf{z}_i[m] - \mathbf{f}_i(\bar{\mathbf{x}}[m])\|^2, \quad (60)$$

where $\mathbf{z}_i[m] \in \mathbb{R}^{M_i}$ is a snapshot of measurements taken at agent i at time m and $\bar{\mathbf{x}}[m]$ is the true state at that time. In this scenario where the measurements stream in sequentially over time, instead of the static scenario elaborated previously, the GGN algorithm can be readily applied to solve such dynamic problems in real-time by initializing $\mathbf{x}_i^0[m]$ with the previous local estimate $\hat{\mathbf{x}}_i[m-1]$. In the following, we show

numerically the applicability of the proposed GGN algorithm on estimating and tracking the state of power systems using real-time power measurements from substations.

VI. NUMERICAL RESULTS

In this section, we compare the GGN algorithm in terms of the objective value (58) and the Mean Square Error (MSE) under the CSE protocol with existing network diffusion algorithms [48] and in particular, we show numerically the applicability of the GGN algorithm to adaptive processing as described in (60) against the method proposed in [24]. The MSE with respect to V_n 's and θ_n 's at the i -th site is

$$\text{MSE}_V^{(i)} = \|\hat{\mathbf{V}}_i - \bar{\mathbf{V}}\|^2, \quad \text{MSE}_\Theta^{(i)} = \|\hat{\boldsymbol{\Theta}}_i - \bar{\boldsymbol{\Theta}}\|^2.$$

We further illustrate the MSE performance of GGN algorithm using the URE protocol in the presence of random link failures. The simulation is based on the IEEE-30 bus ($N = 30$) system in MATPOWER 4.0. The initialization is simply set to 1 for voltage magnitude and 0 for phase.

We take one snapshot of the load profile from the UK National Grid load curve from [55] and scale the base load from MATPOWER on the load buses. Then we run the Optimal Power Flow (OPF) program to determine the generation dispatch for that snapshot. This gives us the true state $\bar{\mathbf{x}}$ and all the quantities $\mathbf{f}(\bar{\mathbf{x}})$, which are all expressed in per unit (p.u.) values. For simplicity we randomly choose 50% of all available measurements $\mathbf{f}(\mathbf{x})$ and generate the measurements $\{\mathbf{z}_i\}_{i=1}^I$ by adding errors $\varepsilon_{i,m} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 10^{-4}$.

A. Comparison with Diffusion Algorithms under CSE Protocol

In this subsection, we evaluate the overall performance of the GGN against existing network diffusion algorithms [48] and its extension to adaptive processing in [24]. To make a fair comparison in terms of communication costs and accuracy, we exploit the coordinated model (i.e., CSE protocol) for [24], [48] and our method, where the agents exchange information synchronously. For simplicity, we divide the entire system into $I = 3$ sites as in Fig. 3. Communication links exist between every two agents $\{i, j\} \in \mathcal{M}$ for $\forall i, j \in \mathcal{I}$, giving an adjacency matrix $\mathbf{A} = \mathbf{1}_I \mathbf{1}_I^T - \mathbf{I}$. The weight matrix used in both cases is constructed according to the Laplacian $\mathbf{L} = \text{diag}(\mathbf{A} \mathbf{1}_I) - \mathbf{A}$ as $\mathbf{W} = \mathbf{I}_I - w \mathbf{L}$ with $w = \beta / \max(\mathbf{A} \mathbf{1}_I)$ and $\beta = 0.3$. The step-size for the GGN algorithm is set as $\alpha_{\text{GGN}} = 0.5$ while $\alpha_{\text{diff}} = 10^{-3}, 10^{-2}, 0.05$ for the sub-gradient methods via network diffusion.

The network diffusion algorithm proceeds as each exchange ℓ takes place, while the GGN algorithm runs $\ell_k = \ell_0 = \ell_* = 3$ exchanges for each k -th update. In particular, the comparison is on the value of the global cost function (58) evaluated using the decentralized estimates

$$\text{Val} = \sum_{i=1}^I \|\mathbf{z}_i - \mathbf{f}_i(\hat{\mathbf{x}}_i)\|^2 \quad (61)$$

and the global phase MSE

$$\text{MSE}_V = \sum_{i=1}^I \text{MSE}_V^{(i)}, \quad \text{MSE}_\Theta = \sum_{i=1}^I \text{MSE}_\Theta^{(i)} \quad (62)$$

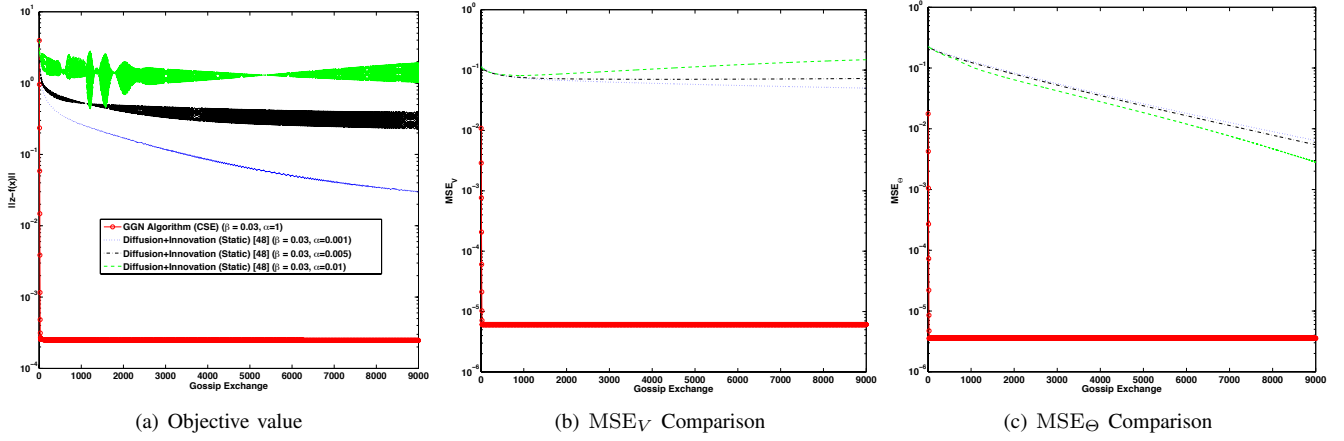


Fig. 4. Comparison between GGN Algorithm (CSE Protocol) and diffusion algorithm in [53] against the gossip exchange with $\ell_* = 3$ exchanges for every update.

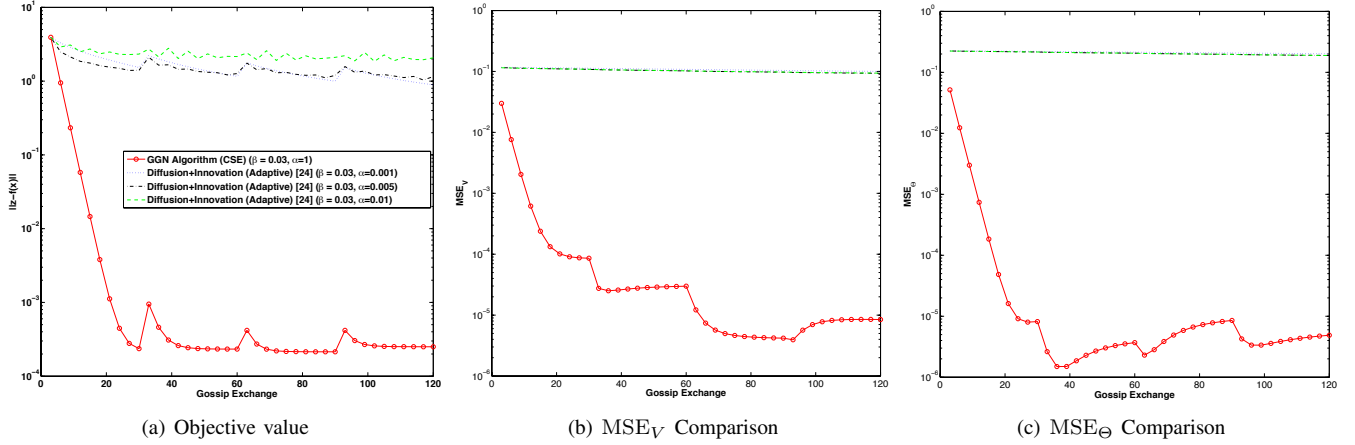


Fig. 5. Comparison between GGN Algorithm (CSE Protocol) and adaptive diffusion algorithm in [24] against the gossip exchange with $\ell_* = 3$ exchanges for every update.

which is plotted against the total number of gossip exchanges so that the comparison is performed on the same time scale. The marker “o” for the GGN algorithm is placed on the scale of each algorithm update k . Since $\ell_k = 3$ for all k , the update-exchange ratio is 1 : 3 such that the marker “o” occurs once every 3 points on the curves.

1) *Estimation on Static Measurements:* In this subsection, we show the comparison between our approach and that in [48] over 9000 gossip exchanges overall. Clearly, the GGN algorithm converges much faster since it reaches the steady state error after $k = 10$ updates (i.e., $k\ell_* = 30$ exchanges). Since the convergence results in Theorem 1 on the bounded perturbation are sufficient but not necessary, it is observed in Fig. 4(a) to 4(c) that the objective values Val and the MSE_Θ of the state estimates still decrease rapidly over time even though the gossip exchange per update $\ell_* = 3$ is small. On the other hand, the objective value and the MSE of the network diffusion algorithm for AC state estimation in [48] decreases slowly until 10000 exchanges with large oscillations (due to constant step-sizes). Network diffusion algorithms using a diminishing step-size $\alpha_{\text{diff}}/\ell$ for $\ell = 1, 2, \dots$ exhibit similar performance except for the oscillations and hence are not presented for clarity. Also, it can be seen that for a non-convex problem, the steady state of the network diffusion algorithm is sensitive

to the choice of the step-size α_{diff} and there exist significant oscillations that persist in the update. On the other hand, the GGN algorithm conditions the gradient by the GN Hessian and therefore the update tends to be smooth and continues to lie in the proximity of the desired solution with high accuracy.

2) *Estimation via Adaptive Processing:* Here we show numerically the applicability of the GGN algorithm to adaptive processing as described in (60) against the method proposed in [24] with the same step-size and network setting. In this simulation, we generate 4 snapshots of measurements $\{\mathbf{z}_i[m]\}_{i=1}^4$ for $m = 1, \dots, 4$ based on the same state $\bar{\mathbf{x}}[m] = \bar{\mathbf{x}}$ by adding i.i.d. Gaussian noise with variance $\sigma^2 = 10^{-4}$, similar to the adaptive setting considered in [24]. More specifically, we use $\ell_* = 3$ gossip exchanges between every two algorithm updates until $k = 10$, thus leading to a total number of 30 exchanges per snapshot. It can be seen from Fig. 5(a) to 5(c) that the proposed GGN algorithm tracks the state estimate accurately when new measurements stream in, where the spikes observed in the plots are caused by the new measurements. Since the number of gossip exchanges is limited, the first order diffusion algorithm in [48] and [24], which suffers from slow convergence, exhibit similar characteristics to those in Fig. 4(a) to 4(c) and fail to track the state as accurately as desired.

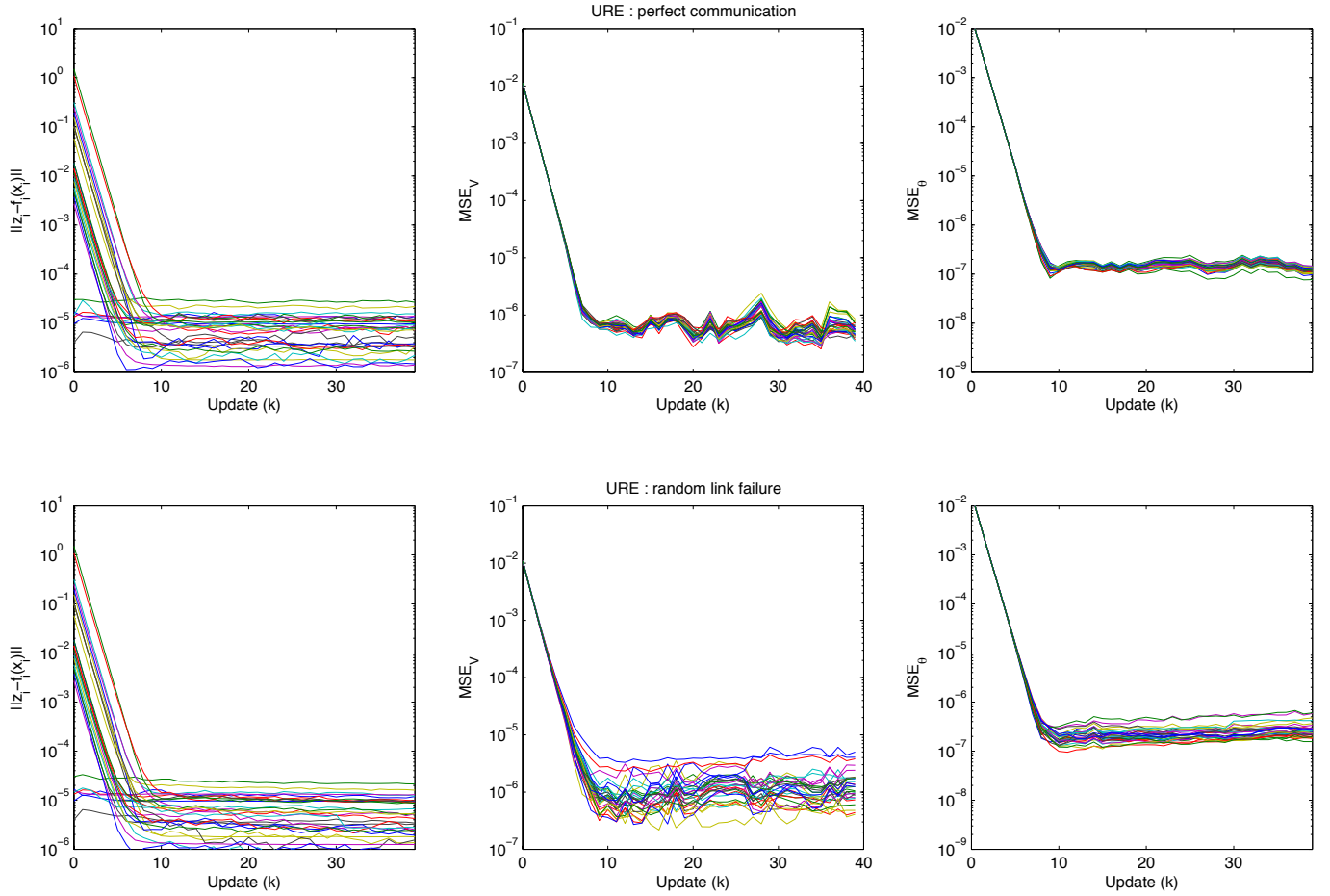


Fig. 6. MSE performance of GGN (URE Protocol) in IEEE-30 bus system with $I = N = 30$ agents and $\mathcal{O}(N)$ pair-wise gossip exchanges. (top) : perfect communication (bottom) : $p = 0.3$ random link failures (each line corresponds to one agent).

B. MSE Performance under URE Protocol with Link Failures

In this section, we examine the MSE performance of the GGN algorithm under the URE protocol with a fixed number of algorithm updates $K = 40$. The performance is evaluated with a demanding setting, where we divide the N -bus system into N sites and each site only communicates with one of its neighbor 10 times on average. The network-wide communication volume in this scenario is on the order of the network diameter $\mathcal{O}(N)$, which implies the number of transmissions in the centralized scheme as if the local measurements are relayed and routed through the entire network. For simplicity, we simulate that at each exchange, the i -th distributed agents wakes up with uniform probability $1/I$ and picks a neighbor with equal probability $1/|\mathcal{M}^{(i)}|$.

In order to show the robustness of the proposed algorithm, we examine the performance of the GGN algorithm for cases with random link failures, where any established link $\{i, j\} \in \mathcal{M}$ fails with probability $p = 0.3$ independently. It is clear that this random communication model with link failures may not satisfy Assumption 1, but it is shown below that our approach is robust to the random setting and degrades gracefully with the probability of link failures. As performance benchmarks,

we track both the individual cost function

$$\text{Val}^{(i)} = \|\mathbf{z}_i - \mathbf{f}_i(\hat{\mathbf{x}}_i)\|$$

evaluated by local estimates $\hat{\mathbf{x}}_i$ and the individual voltage and phase MSE of the estimation $\text{MSE}_V^{(i)}$ and $\text{MSE}_\theta^{(i)}$ in Fig. 6. It can be observed from the figure that the MSE curves of state estimates of different sites are highly consistent and they all converge asymptotically when there is no link failures. Similar behaviors can be observed for the case with random link failures, where the local estimate at each site is not in perfect consistence with the others, but the accuracy remains satisfactory compared to the perfect case and degrades gracefully with the probability of link failures.

VII. CONCLUSIONS

In this paper, we study the convergence and performance of a Gossip-based Gauss-Newton (GGN) algorithm, which can serve as a guideline in solving similar problems. We also demonstrate its application in power system state estimation, where numerical results suggest that the proposed algorithm leads to accurate state estimates across the distributed areas, which greatly improves the robustness against link/node failures or network congestions with communication and computations that scale similarly to the centralized scheme.

APPENDIX A
PROOF OF LEMMA 2

To study the convergence of the GGN algorithm, we examine the distributed update

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}} \left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k(\ell_k) \right], \quad (63)$$

which can be written with respect to the descent \mathbf{d}_i^k in (11)

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}} \left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k + \alpha \left(\mathbf{d}_i^k - \mathbf{d}_i^k(\ell_k) \right) \right]. \quad (64)$$

By subtracting the fixed point \mathbf{x}^* and using the non-expansive property of the operator $P_{\mathbb{X}}(\cdot)$ on the closed convex set \mathbb{X} , we have the following recursion

$$\|\mathbf{x}_i^{k+1} - \mathbf{x}^*\| \leq \|\mathbf{x}_i^k - \mathbf{x}^* - \alpha \mathbf{d}_i^k\| + \alpha \|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|.$$

For convenience, we denote $\mathbf{G}^\dagger(\cdot)$ as the pseudo-inverse of $\mathbf{G}(\cdot)$. For any fixed point $\mathbf{x}^* \in \mathbb{X}$ in (4) such that $\mathbf{G}^\dagger(\mathbf{x}^*)\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$, the first term can be equivalently written by substituting (8) as follows

$$\begin{aligned} \mathbf{x}_i^k - \mathbf{x}^* - \alpha \mathbf{d}_i^k &= \mathbf{x}_i^k - \mathbf{x}^* \\ &\quad - \alpha \mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k) + \alpha \mathbf{G}^\dagger(\mathbf{x}^*)\mathbf{g}(\mathbf{x}^*). \end{aligned} \quad (65)$$

Using (3) together with the invertibility condition of $\mathbf{G}(\mathbf{x})$ over $\mathbf{x} \in \mathbb{X}$ in Assumption 3, we have

$$\mathbf{x}_i^k - \mathbf{x}^* = \mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k) (\mathbf{x}_i^k - \mathbf{x}^*). \quad (66)$$

Then by substituting (66) into (65), and meanwhile adding and subtracting simultaneously a term $\alpha \mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}^*)$, we have the following expression

$$\begin{aligned} \mathbf{x}_i^k - \mathbf{x}^* - \alpha \mathbf{d}_i^k &= \mathbf{G}^\dagger(\mathbf{x}_i^k) \left[\mathbf{G}(\mathbf{x}_i^k) (\mathbf{x}_i^k - \mathbf{x}^*) - \alpha \mathbf{g}(\mathbf{x}_i^k) + \alpha \mathbf{g}(\mathbf{x}^*) \right] \\ &\quad + \alpha \left[\mathbf{G}^\dagger(\mathbf{x}^*) - \mathbf{G}^\dagger(\mathbf{x}_i^k) \right] \mathbf{g}(\mathbf{x}^*). \end{aligned}$$

The expression in the first term can be re-written with the mean-value theorem as follows

$$\begin{aligned} \alpha \mathbf{g}(\mathbf{x}^*) - \alpha \mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{x}^* - \mathbf{x}) & \quad (67) \\ &= \alpha \left[\int_0^1 \mathbf{G}(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x}))(\mathbf{x}^* - \mathbf{x}) dt \right] - \mathbf{G}(\mathbf{x})(\mathbf{x}^* - \mathbf{x}) \\ &= \alpha \left(\int_0^1 [\mathbf{G}(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - \mathbf{G}(\mathbf{x})] (\mathbf{x}^* - \mathbf{x}) dt \right) \\ &\quad - (1 - \alpha) \mathbf{G}(\mathbf{x})(\mathbf{x}^* - \mathbf{x}), \end{aligned}$$

whose norm can be bounded by using Assumption 3 as

$$\begin{aligned} \|\alpha \mathbf{g}(\mathbf{x}^*) - \alpha \mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{x}^* - \mathbf{x})\| & \quad (68) \\ &\leq \alpha \left[\int_0^1 \|\mathbf{G}(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - \mathbf{G}(\mathbf{x})\| dt \right] \|\mathbf{x} - \mathbf{x}^*\| \\ &\quad + (1 - \alpha) \sigma_{\max} \|\mathbf{x} - \mathbf{x}^*\|. \end{aligned}$$

From the Lipschitz condition in Assumption 3, we have

$$\int_0^1 \|\mathbf{G}(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - \mathbf{G}(\mathbf{x})\| dt \leq \omega \|\mathbf{x} - \mathbf{x}^*\| \int_0^1 t dt.$$

Thus, if condition (3) of Assumption 3 holds, we have

$$\begin{aligned} \|\alpha \mathbf{g}(\mathbf{x}^*) - \alpha \mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{x}^* - \mathbf{x})\| \\ \leq \frac{\omega}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 + (1 - \alpha) \sigma_{\max} \|\mathbf{x} - \mathbf{x}^*\| \end{aligned}$$

and finally according to [56, Lemma 1], we have

$$\begin{aligned} \|\mathbf{G}^\dagger(\mathbf{x}) - \mathbf{G}^\dagger(\mathbf{x}^*)\| \\ \leq \sqrt{2} \|\mathbf{G}^\dagger(\mathbf{x})\| \|\mathbf{G}^\dagger(\mathbf{x}^*)\| \|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}^*)\| \\ \leq \frac{\sqrt{2}\omega}{\sigma_{\min}^2} \|\mathbf{x} - \mathbf{x}^*\|. \end{aligned}$$

By the definition of matrix norm, we have $\|\mathbf{G}^\dagger(\mathbf{x})\|^2 = \|\mathbf{G}^\dagger(\mathbf{x}) (\mathbf{G}^\dagger(\mathbf{x}))^T\| = \|(\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x}))^{-1}\|$. Then evidently, the norm of a matrix inverse corresponds to the reciprocal of the eigenvalue of that matrix with the smallest magnitude. By Assumption 3 condition (2), this quantity is bounded as σ_{\min}^2 , and therefore $\|\mathbf{G}^\dagger(\mathbf{x})\| \leq 1/\sigma_{\min}$. Thus finally, using $\|\mathbf{G}^\dagger(\mathbf{x})\| \leq 1/\sigma_{\min}$ and the above inequalities gives the following bounds

$$\|\mathbf{x}_i^k - \mathbf{x}^* - \alpha \mathbf{d}_i^k\| \leq T_1 \|\mathbf{x}_i^k - \mathbf{x}^*\|^2 + T_2 \|\mathbf{x}_i^k - \mathbf{x}^*\|,$$

where $\epsilon_* \triangleq \|\mathbf{g}(\mathbf{x}^*)\|$ indicates the goodness of fit at \mathbf{x}^* and

$$T_1 \triangleq \frac{\omega}{2\sigma_{\min}}, \quad T_2 \triangleq (1 - \alpha) \frac{\sigma_{\max}}{\sigma_{\min}} + \alpha \frac{\sqrt{2}\omega\epsilon_*}{\sigma_{\min}^2}. \quad (69)$$

Therefore, we have the error recursion (31).

APPENDIX B
PROOF OF LEMMA 3

Using (21), we evaluate the deviation of $\phi_k(\ell)$ from the average $\bar{\phi}_k = [\mathbf{1}^T \otimes \mathbf{I}_{N_\phi}] \phi_k(0)/I$ for a finite ℓ . By subtracting the average $\bar{\phi}_k$ on both sides of (21), we have

$$\begin{aligned} \phi_k(\ell) - \bar{\phi}_k &= [\mathbf{W}_k(\ell) \otimes \mathbf{I}_{N_\phi}] \phi_k(\ell - 1) - \frac{\mathbf{1}^T \otimes \mathbf{I}_{N_\phi}}{I} \phi_k(0) \\ &= \left[\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') \otimes \mathbf{I}_{N_\phi} \right] \phi_k(0) - \frac{\mathbf{1}^T \otimes \mathbf{I}_{N_\phi}}{I} \phi_k(0) \\ &= \left[\left(\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}^T}{I} \right) \otimes \mathbf{I}_{N_\phi} \right] \phi_k(0). \end{aligned}$$

Then, we bound the norms of the above equation as

$$\|\phi_k(\ell) - \bar{\phi}_k\| \leq \left\| \prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}^T}{I} \right\| \|\phi_k(0)\|. \quad (70)$$

Using Lemma 1 and the norm inequality $\|\cdot\| \leq \|\cdot\|_F$, we have

$$\begin{aligned} \|\phi_k(\ell) - \bar{\phi}_k\| &\leq \left\| \prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}^T}{I} \right\|_F \|\phi_k(0)\| \\ &\leq \left[2I^2 \frac{1 + \eta^{-L_0}}{1 - \eta^{L_0}} (1 - \eta^{L_0})^{\ell/L_0} \right] \|\phi_k(0)\|. \end{aligned}$$

The quantity $\|\phi_k(0)\|$ is by definition (14) determined as

$$\|\phi_k(0)\|^2 = \sum_{i=1}^I \|\mathbf{h}_{k,i}(0)\|^2 + \sum_{i=1}^I \|\mathbf{H}_{k,i}(0)\|_F^2 \quad (71)$$

$$= \sum_{i=1}^I \|\mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{g}_i(\mathbf{x}_i^k)\|^2 + \sum_{i=1}^I \|\mathbf{G}_i^T(\mathbf{x}_i^k) \mathbf{G}_i(\mathbf{x}_i^k)\|_F^2 \\ \leq I\sigma_{\max}^2(C_g^2 + N\sigma_{\max}^2), \quad (72)$$

where the norm inequality is used $\|\mathbf{G}_i^T(\mathbf{x}) \mathbf{G}_i(\mathbf{x})\|_F^2 \leq \|\mathbf{G}^T(\mathbf{x}) \mathbf{G}(\mathbf{x})\|_F^2 \leq N \|\mathbf{G}^T(\mathbf{x}) \mathbf{G}(\mathbf{x})\|_2^2 = N\sigma_{\max}^4$. Letting $C = 2I^2 \sqrt{I\sigma_{\max}^2(C_g^2 + N\sigma_{\max}^2)(1 + \eta^{-L_0})/(1 - \eta^{L_0})}$ and $\lambda_\eta = (1 - \eta^{L_0})^{1/L_0}$, then the error is bounded as

$$\|\phi_k(\ell) - \bar{\phi}_k\| \leq C\lambda_\eta^\ell. \quad (73)$$

By definition of the errors $\mathbf{e}_{k,i}(\ell)$ and $\mathbf{E}_{k,i}(\ell)$, we have

$$\phi_k(\ell) - \bar{\phi}_k = \begin{bmatrix} \mathbf{e}_{k,1}(\ell) \\ \text{vec}[\mathbf{E}_{k,1}(\ell)] \\ \vdots \\ \mathbf{e}_{k,I}(\ell) \\ \text{vec}[\mathbf{E}_{k,I}(\ell)] \end{bmatrix}, \quad (74)$$

and therefore the norm of each component in the error vector $\phi_k(\ell) - \bar{\phi}_k$ is bounded by the full norm

$$\|\mathbf{e}_k(\ell)\| < C\lambda_\eta^\ell, \quad \|\mathbf{E}_k(\ell)\|_F < C\lambda_\eta^\ell. \quad (75)$$

APPENDIX C PROOF OF LEMMA 4

We prove this result by mathematical induction. During the proof, we will repetitively use the result in [52] that for any given matrix \mathbf{Z} and $\delta\mathbf{Z}$, the matrix series expansion

$$(\mathbf{Z} + \delta\mathbf{Z})^{-1} = \sum_{q=0}^{\infty} (-1)^{q+1} (\mathbf{Z}^{-1}\delta\mathbf{Z})^q \mathbf{Z}^{-1} \quad (76)$$

holds as long as $\|\mathbf{Z}^{-1}\delta\mathbf{Z}\| < 1$.

A. Initial Case: $k = 1$

When $k = 1$ and given $\mathbf{x}_i^0 = \mathbf{x}^0$ for all i , we have

$$\|\mathbf{x}_i^1 - \mathbf{x}_j^1\| \leq \|\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0)\|, \quad (77)$$

where the discrepancy is expressed explicitly as

$$\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0) = [\bar{\mathbf{H}}_0 + \mathbf{E}_{0,i}(\ell_0)]^{-1} [\bar{\mathbf{h}}_0 + \mathbf{e}_{0,i}(\ell_0)] \\ - [\bar{\mathbf{H}}_0 + \mathbf{E}_{0,j}(\ell_0)]^{-1} [\bar{\mathbf{h}}_0 + \mathbf{e}_{0,j}(\ell_0)]. \quad (78)$$

Thus, if the perturbations $\mathbf{E}_{0,i}(\ell_0)$, $\mathbf{E}_{0,j}(\ell_0)$ are small enough, the expansion in (76) can be applied here to simplify the expression.

1) **Matrix series expansion:** Since $\mathbf{x}_i^0 = \mathbf{x}^0$ for all i such that $\bar{\mathbf{H}}_0 = \mathbf{R}(\mathbf{x}_i^0)$ and $\bar{\mathbf{h}}_0 = \mathbf{r}(\mathbf{x}_i^0)$, they can be bounded as

$$\|\bar{\mathbf{h}}_0\| = \|\mathbf{r}(\mathbf{x}_i^0)\| \leq \frac{\sigma_{\max} C_g}{I}, \quad (79)$$

$$\|\bar{\mathbf{H}}_0^{-1}\| = \|\mathbf{R}^{-1}(\mathbf{x}_i^0)\| \leq \frac{I}{\sigma_{\min}^2}. \quad (80)$$

Note that from the norm equality of sub-matrices

$$\|\bar{\mathbf{H}}_0^{-1} \mathbf{E}_{0,i}(\ell_0)\| \leq \|\bar{\mathbf{H}}_0^{-1}\| \|\mathbf{E}_{0,i}(\ell_0)\| \leq \|\bar{\mathbf{H}}_0^{-1}\| \|\mathbf{E}_0(\ell_0)\|_F \\ \|\bar{\mathbf{H}}_0^{-1} \mathbf{E}_{0,j}(\ell_0)\| \leq \|\bar{\mathbf{H}}_0^{-1}\| \|\mathbf{E}_{0,j}(\ell_0)\| \leq \|\bar{\mathbf{H}}_0^{-1}\| \|\mathbf{E}_0(\ell_0)\|_F,$$

and by Assumption 4 we have $\ell_0 \geq \ell_*$. From Lemma 3 and again Assumption 4, the above upper bound can be further bounded as

$$\|\bar{\mathbf{H}}_0^{-1}\| \|\mathbf{E}_0(\ell_0)\|_F \leq \frac{I}{\sigma_{\min}^2} C\lambda_\eta^{\ell_0} \leq \frac{I}{\sigma_{\min}^2} \frac{C\kappa}{CC_\sigma(1 + DC_\sigma)} \\ = \frac{I/\sigma_{\min}^2}{C_\sigma} \frac{1}{1 + DC_\sigma} \kappa < \kappa \ll 1.$$

Therefore, letting $\delta\mathbf{Z} = \mathbf{E}_{0,i}(\ell_0)$ or $\mathbf{E}_{0,j}(\ell_0)$ and $\mathbf{Z} = \bar{\mathbf{H}}_0$, then it follows that the expansion holds. By grouping all the high order terms $q \geq 2$ and bound them with $\mathcal{O}(\kappa^2)$, we have

$$\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0) = \bar{\mathbf{H}}_0^{-1} [\mathbf{E}_{0,i}(\ell_0) - \mathbf{E}_{0,j}(\ell_0)] \bar{\mathbf{H}}_0^{-1} \bar{\mathbf{h}}_0 \\ - \bar{\mathbf{H}}_0^{-1} [\mathbf{e}_{0,i}(\ell_0) - \mathbf{e}_{0,j}(\ell_0)] + \mathcal{O}(\kappa^2). \quad (81)$$

2) **Proof of success when $k = 1$:** Similarly by Lemma 3, we can bound

$$\|\mathbf{E}_{0,i}(\ell_0) - \mathbf{E}_{0,j}(\ell_0)\| \leq 2C\lambda_\eta^{\ell_0} = \mathcal{O}(\kappa), \\ \|\mathbf{e}_{0,i}(\ell_0) - \mathbf{e}_{0,j}(\ell_0)\| \leq 2C\lambda_\eta^{\ell_0} = \mathcal{O}(\kappa).$$

From (50) and (81), we have

$$\|\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0)\| \\ \leq C\lambda_\eta^{\ell_0} \cdot 2 \left(\|\bar{\mathbf{H}}_0^{-1}\| + \|\bar{\mathbf{H}}_0^{-1}\|^2 \|\bar{\mathbf{h}}_0\| \right) + \mathcal{O}(\kappa^2) \\ \leq C\lambda_\eta^{\ell_0} \cdot 2 \left(\frac{\sigma_{\max} C_g I}{\sigma_{\min}^4} + \frac{I}{\sigma_{\min}^2} \right) + \mathcal{O}(\kappa^2) \\ \leq CC_\sigma \lambda_\eta^{\ell_*} D_0 + \mathcal{O}(\kappa^2), \text{ where } D_0 = \lambda_\eta^{(\ell_0 - \ell_*)}.$$

Ignoring high order terms $\mathcal{O}(\kappa^2)$, we have

$$\|\mathbf{x}_i^1 - \mathbf{x}_j^1\| \leq CC_\sigma \lambda_\eta^{\ell_*} D_0 = \mathcal{O}(\kappa) \quad (82)$$

and therefore the result holds for $k = 1$.

B. Induction: $k = K$ and $k = K + 1$

Let the error bound holds for $k = K$ such that

$$\|\mathbf{x}_i^K - \mathbf{x}_j^K\| \leq CC_\sigma \lambda_\eta^{\ell_*} D_{K-1}. \quad (83)$$

Similarly, the triangle inequality holds

$$\|\mathbf{x}_i^{K+1} - \mathbf{x}_j^{K+1}\| \leq \|\mathbf{x}_i^K - \mathbf{x}_j^K\| + \|\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)\|,$$

where

$$\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K) \\ = [\bar{\mathbf{H}}_K + \mathbf{E}_{K,i}(\ell_K)]^{-1} [\bar{\mathbf{h}}_K + \mathbf{e}_{K,i}(\ell_K)] \\ - [\bar{\mathbf{H}}_K + \mathbf{E}_{K,j}(\ell_K)]^{-1} [\bar{\mathbf{h}}_K + \mathbf{e}_{K,j}(\ell_K)]. \quad (84)$$

Similar to the case when $k = 1$, if the perturbations $\mathbf{E}_{K,i}(\ell_K)$, $\mathbf{E}_{K,j}(\ell_K)$ are small enough, the expansion in (76) can be applied here to simplify the expression.

1) **Matrix series expansion:** By definition (45), we have

$$\|\bar{\mathbf{H}}_K^{-1}\| = \left\| [\mathbf{R}(\mathbf{x}_i^K) + \mathbf{\Delta}_{K,i}]^{-1} \right\|, \quad (85)$$

which is another perturbed inverse. Thus we first examine whether this inverse can be expanded using the series expansion in (76). From (46) and (83), we have

$$\|\Delta_{K,i}\| \leq \nu_\Delta CC_\sigma \lambda_\eta^{\ell_*} D_{K-1} < \nu_\Delta CC_\sigma \lambda_\eta^{\ell_*} D, \quad (86)$$

where the last inequality comes from the non-negativity of λ_η (i.e., $D > D_k$ for all finite k). By Assumption 4, we have

$$\begin{aligned} \|\mathbf{R}^{-1}(\mathbf{x}_i^K) \Delta_{K,i}\| &\leq \|\mathbf{R}^{-1}(\mathbf{x}_i^K)\| \|\Delta_{K,i}\| \\ &\leq \frac{I}{\sigma_{\min}^2} \frac{\nu_\Delta CC_\sigma D \kappa}{CC_\sigma(1 + DC_\sigma)} \kappa \\ &< \frac{\nu_\Delta I / \sigma_{\min}^2}{C_\sigma} \frac{DC_\sigma}{1 + DC_\sigma} \kappa < \kappa. \end{aligned} \quad (87)$$

Therefore, the matrix series expansion holds for (85). Then by grouping all the high order terms and bound them with $\mathcal{O}(\kappa^2)$, using the above calculations we have

$$\begin{aligned} \|\bar{\mathbf{H}}_K^{-1}\| &\leq \|\mathbf{R}^{-1}(\mathbf{x}_i^K)\| + \|\mathbf{R}^{-1}(\mathbf{x}_i^K)\|^2 \|\Delta_{K,i}\| + \mathcal{O}(\kappa^2) \\ &\leq \frac{I}{\sigma_{\min}^2} + \frac{I}{\sigma_{\min}^2} \kappa + \mathcal{O}(\kappa^2). \end{aligned} \quad (88)$$

With (88), we have

$$\begin{aligned} \|\bar{\mathbf{H}}_K^{-1} \mathbf{E}_{K,i}(\ell_K)\| &\leq \|\bar{\mathbf{H}}_K^{-1}\| \|\mathbf{E}_{K,i}(\ell_K)\| \leq \|\bar{\mathbf{H}}_K^{-1}\| \|\mathbf{E}_K(\ell_K)\|_F \\ \|\bar{\mathbf{H}}_K^{-1} \mathbf{E}_{K,j}(\ell_K)\| &\leq \|\bar{\mathbf{H}}_K^{-1}\| \|\mathbf{E}_{K,j}(\ell_K)\| \leq \|\bar{\mathbf{H}}_K^{-1}\| \|\mathbf{E}_K(\ell_K)\|_F, \end{aligned}$$

and there is $\ell_K \geq \ell_*$. Then by Lemma 3 and again Assumption 4, the above upper bound can be further bounded as

$$\begin{aligned} \|\bar{\mathbf{H}}_K^{-1}\| \|\mathbf{E}_K(\ell_K)\|_F &\leq \left(\frac{I}{\sigma_{\min}^2} + \frac{I}{\sigma_{\min}^2} \kappa + \mathcal{O}(\kappa^2) \right) C \lambda_\eta^{\ell_K} \\ &< \kappa + \mathcal{O}(\kappa^2) \ll 1, \end{aligned}$$

thus the matrix series expansion also holds for (84). Then by grouping all the higher order terms and bound them with $\mathcal{O}(\kappa^2)$, we have

$$\begin{aligned} \mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K) &= \bar{\mathbf{H}}_K^{-1} [\mathbf{E}_{K,i}(\ell_K) - \mathbf{E}_{K,j}(\ell_K)] \bar{\mathbf{H}}_K^{-1} \bar{\mathbf{h}}_K \\ &\quad - \bar{\mathbf{H}}_K^{-1} [\mathbf{e}_{K,i}(\ell_K) - \mathbf{e}_{K,j}(\ell_K)] + \mathcal{O}(\kappa^2). \end{aligned}$$

2) Proof of success when $k = K + 1$: From Lemma 3, we can bound

$$\begin{aligned} \|\mathbf{E}_{K,i}(\ell_K) - \mathbf{E}_{K,j}(\ell_K)\| &\leq 2C \lambda_\eta^{\ell_K} = \mathcal{O}(\kappa), \\ \|\mathbf{e}_{K,i}(\ell_K) - \mathbf{e}_{K,j}(\ell_K)\| &\leq 2C \lambda_\eta^{\ell_K} = \mathcal{O}(\kappa) \end{aligned}$$

and accordingly

$$\begin{aligned} \|\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)\| &\leq C \lambda_\eta^{\ell_K} 2 \left(\underbrace{\|\bar{\mathbf{H}}_K^{-1}\| + \|\bar{\mathbf{H}}_K^{-1}\|^2 \|\bar{\mathbf{h}}_K\|}_{\leq C_\sigma} \right) + \mathcal{O}(\kappa^2) \\ &\leq CC_\sigma \lambda_\eta^{\ell_K} + \mathcal{O}(\kappa^2). \end{aligned}$$

Finally by ignoring high order terms $\mathcal{O}(\kappa^2)$, we have

$$\begin{aligned} \|\mathbf{x}_i^{K+1} - \mathbf{x}_j^{K+1}\| &\leq \|\mathbf{x}_i^K - \mathbf{x}_j^K\| + CC_\sigma \lambda_\eta^{\ell_K} \\ &\leq CC_\sigma \lambda_\eta^{\ell_*} \left(D_{K-1} + \lambda_\eta^{(\ell_K - \ell_*)} \right) \\ &= CC_\sigma \lambda_\eta^{\ell_*} D_K, \end{aligned}$$

and therefore the result holds for $k = K + 1$ given $k = K$.

APPENDIX D PROOF OF PROPOSITION 1

By the decomposition in (48), we have

$$\begin{aligned} \mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k &= [\mathbf{R}(\mathbf{x}_i^k) + \Delta_{k,i} + \mathbf{E}_{k,i}(\ell_k)]^{-1} [\mathbf{r}(\mathbf{x}_i^k) + \delta_{k,i} + \mathbf{e}_{k,i}(\ell_k)] \\ &\quad - \mathbf{R}(\mathbf{x}_i^k)^{-1} \mathbf{r}(\mathbf{x}_i^k). \end{aligned} \quad (89)$$

Now that we verify that the matrix series expansion holds for similar approximations. First of all, from Lemma 4 and (46), we have $\|\Delta_{k,i}(\ell_k)\| \leq \nu_\Delta CC_\sigma \lambda_\eta^{\ell_*} D_{k-1}$. The expansion depends on the quantity

$$\begin{aligned} \|\mathbf{R}^{-1}(\mathbf{x}_i^k) (\Delta_{k,i} + \mathbf{E}_{k,i}(\ell_k))\| &\leq \|\mathbf{R}^{-1}(\mathbf{x}_i^k)\| \|\Delta_{k,i}\| + \|\mathbf{R}^{-1}(\mathbf{x}_i^k)\| \|\mathbf{E}_{k,i}(\ell_k)\|. \end{aligned}$$

Using the derivation in (87), we have

$$\begin{aligned} \|\mathbf{R}^{-1}(\mathbf{x}_i^k) (\Delta_{k,i} + \mathbf{E}_{k,i}(\ell_k))\| &< \kappa + \frac{I}{\sigma_{\min}^2} C \lambda_\eta^{\ell_k} < \kappa + \frac{I / \sigma_{\min}^2}{C_\sigma} \frac{1}{(1 + DC_\sigma)} \kappa < 2\kappa \ll 1, \end{aligned}$$

and the expansion holds. The difference can then be written by grouping the high order terms $\mathcal{O}(\kappa^2)$ as

$$\begin{aligned} \mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k &= -\mathbf{R}^{-1}(\mathbf{x}_i^k) [\Delta_{k,i} + \mathbf{E}_{k,i}(\ell_k)] \mathbf{R}^{-1}(\mathbf{x}_i^k) \mathbf{r}(\mathbf{x}_i^k) \\ &\quad + \mathbf{R}^{-1}(\mathbf{x}_i^k) [\delta_{k,i} + \mathbf{e}_{k,i}(\ell_k)] + \mathcal{O}(\epsilon^2). \end{aligned}$$

Note that $\|\delta_{k,i}(\ell_k)\| \leq \nu_\delta CC_\sigma \lambda_\eta^{\ell_*} D_{k-1}$, we have

$$\begin{aligned} \|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| &\leq (\nu_\Delta CC_\sigma \lambda_\eta^{\ell_*} D_{k-1} + C \lambda_\eta^{\ell_k}) \frac{\sigma_{\max} C_g I}{\sigma_{\min}^4} \\ &\quad + (\nu_\delta CC_\sigma \lambda_\eta^{\ell_*} D_{k-1} + C \lambda_\eta^{\ell_k}) \frac{I}{\sigma_{\min}^2} \\ &\leq CC_\sigma D_{k-1} \left(\frac{\nu_\Delta \sigma_{\max} C_g I}{\sigma_{\min}^4} + \frac{\nu_\delta I}{\sigma_{\min}^2} \right) \lambda_\eta^{\ell_*} \\ &\quad + C \left(\frac{\sigma_{\max} C_g I}{\sigma_{\min}^4} + \frac{I}{\sigma_{\min}^2} \right) \lambda_\eta^{\ell_*}. \end{aligned}$$

Using (50) to bound the above inequality, we have

$$\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| \leq CC_\sigma (1 + D_{k-1} C_\sigma) \lambda_\eta^{\ell_*}. \quad (90)$$

Considering that $D \geq D_{k-1}$ for any k , then using Assumption 4 gives us the following bound

$$\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| < \kappa \quad (91)$$

for all i and k .

REFERENCES

- [1] J. Nocedal and S. Wright, *Numerical Optimization*. Springer verlag, 1999.
- [2] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial Mathematics, 1996, vol. 16.
- [3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ Pr, 2004.
- [4] Å. Björck, *Numerical Methods for Least Squares Problems*. Society for Industrial Mathematics, 1996, no. 51.

- [5] A. Monticelli, "Electric Power System State Estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.
- [6] C. Mensing and S. Plass, "Positioning Algorithms for Cellular Networks using TDOA," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4. IEEE, 2006, pp. IV–IV.
- [7] P. Stoica, R. Moses, B. Friedlander, and T. Soderstrom, "Maximum Likelihood Estimation of the Parameters of Multiple Sinusoids from Noisy Measurements," *Acoustics, Speech and IEEE Trans. Signal Process.*, vol. 37, no. 3, pp. 378–392, 1989.
- [8] B. Bell and F. Cathey, "The Iterated Kalman Filter Update as a Gauss-Newton Method," *Automatic Control, IEEE Transactions on*, vol. 38, no. 2, pp. 294–297, 1993.
- [9] M. Schweiger, S. Arridge, and I. Nissilä, "Gauss-Newton method for Image Reconstruction in Diffuse Optical Tomography," *Physics in medicine and biology*, vol. 50, p. 2365, 2005.
- [10] J. Tsitsiklis, "Problems in Decentralized Decision Making and Computation," DTIC Document, Tech. Rep., 1984.
- [11] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking, "Randomized Rumor Spreading," in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 2000, pp. 565–574.
- [12] R. Olfati-Saber and R. Murray, "Consensus Problems in Networks of Agents with Switching Topology and Time-Delays," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1520–1533, 2004.
- [13] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip Algorithms for Distributed Signal Processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.
- [14] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based Computation of Aggregate Information," in *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*. IEEE, 2003, pp. 482–491.
- [15] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.
- [16] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and Cooperation in Networked Multi-agent Systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [17] B. Johansson, T. Keviczky, M. Johansson, and K. Johansson, "Subgradient Methods and Consensus Algorithms for Solving Convex Optimization Problems," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*. IEEE, 2008, pp. 4185–4190.
- [18] D. Bertsekas, M. I. of Technology, Laboratory for Information, and D. Systems, "A New Class of Incremental Gradient Methods for Least Squares Problems," *SIAM Journal on Optimization*, vol. 7, no. 4, pp. 913–926, 1997.
- [19] A. Nedic and D. Bertsekas, "Incremental Subgradient Methods for Non-differentiable Optimization," *SIAM Journal of Optimization*, vol. 12, no. 1, pp. 109–138, 2001.
- [20] A. Nedic and A. Ozdaglar, "Distributed Subgradient Methods for Multi-agent Optimization," *Automatic Control, IEEE Transactions on*, vol. 54, no. 1, pp. 48–61, 2009.
- [21] S. Ram, A. Nedic, and V. Veeravalli, "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.
- [22] A. Nedic, "Asynchronous Broadcast-based Convex Optimization over a Network," *Automatic Control, IEEE Transactions on*, no. 99, pp. 1–1, 2010.
- [23] K. Srivastava and A. Nedic, "Distributed Asynchronous Constrained Stochastic Optimization," *Selected Topics in Signal Processing, IEEE Journal of*, no. 99, pp. 1–1, 2011.
- [24] S. Kar, J. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *Information Theory, IEEE Transactions on*, vol. 58, no. 6, pp. 3575–3605, june 2012.
- [25] I. Matei and J. Baras, "Performance Evaluation of the Consensus-based Distributed Subgradient Method under Random Communication Topologies," *Selected Topics in Signal Processing, IEEE Journal of*, no. 99, pp. 1–1, 2011.
- [26] J. Chen and A. Sayed, "Diffusion Adaptation Strategies for Distributed Optimization and Learning over Networks," *Signal Processing, IEEE Transactions on*, vol. 60, no. 8, pp. 4289–4305, 2012.
- [27] C. Lopes and A. Sayed, "Diffusion Least-Mean Squares over Adaptive Networks: Formulation and Performance Analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, 2008.
- [28] F. Cattivelli and A. Sayed, "Diffusion LMS Strategies for Distributed Estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, 2010.
- [29] F. Cattivelli, C. Lopes, and A. Sayed, "Diffusion Recursive Least-Squares for Distributed Estimation over Adaptive Networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, 2008.
- [30] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A Distributed Newton Method for Network Utility Maximization," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 1816–1821.
- [31] M. Ilic and A. Hsu, "Toward Distributed Contingency Screening using Line Flow Calculators and Dynamic Line Rating Units (DLRs)," in *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2012, pp. 2027–2035.
- [32] B. Bejar, P. Belanovic, and S. Zazo, "Distributed Gauss-Newton Method for Localization in Ad-hoc Networks," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 1452–1454.
- [33] B. Cheng, R. Hudson, F. Lorenzelli, L. Vandenbergh, and K. Yao, "Distributed Gauss-Newton Method for Node Localization in Wireless Sensor Networks," in *Signal Processing Advances in Wireless Communications, 2005 IEEE 6th Workshop on*. IEEE, 2005, pp. 915–919.
- [34] G. Calafiore, L. Carlone, and M. Wei, "A Distributed Gauss-Newton Approach for Range-based Localization of Multi-agent Formations," in *Computer-Aided Control System Design (CACSD), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1152–1157.
- [35] T. Zhao and A. Nehorai, "Information-Driven Distributed Maximum Likelihood Estimation based on Gauss-Newton Method in Wireless Sensor Networks," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4669–4682, 2007.
- [36] F. Schweppe and E. Handschin, "Static State Estimation in Electric Power Systems," *Proceedings of the IEEE*, vol. 62, no. 7, pp. 972–982, 1974.
- [37] R. Larson, W. Tinney, and J. Peschon, "State Estimation in Power Systems Part I: Theory and Feasibility," *IEEE Trans. Power App. Syst.*, no. 3Part-I, pp. 345–352, 1970.
- [38] C. Brice and R. Cavin, "Multiprocessor Static State Estimation," *IEEE Trans. Power App. Syst.*, no. 2, pp. 302–308, 1982.
- [39] M. Kurzyn, "Real-Time State Estimation for Large-Scale Power Systems," *IEEE Trans. Power App. Syst.*, no. 7, pp. 2055–2063, 1983.
- [40] T. Yang, H. Sun, and A. Bose, "Transition to a Two-Level Linear State Estimator : Part i & ii," *IEEE Trans. Power Syst.*, no. 99, pp. 1–1, 2011.
- [41] A. Gómez-Expósito, A. Abur, A. de la Villa Jaén, and C. Gómez-Quiles, "A Multilevel State Estimation Paradigm for Smart Grids," *Proceedings of the IEEE*, no. 99, pp. 1–25, 2011.
- [42] D. Falcao, F. Wu, and L. Murphy, "Parallel and Distributed State Estimation," *IEEE Trans. Power Syst.*, vol. 10, no. 2, pp. 724–730, 1995.
- [43] S. Lin, "A Distributed State Estimator for Electric Power Systems," *IEEE Trans. Power Syst.*, vol. 7, no. 2, pp. 551–557, 1992.
- [44] R. Ebrahimi and R. Baldick, "State Estimation Distributed Processing," *IEEE Trans. Power Syst.*, vol. 15, no. 4, pp. 1240–1246, 2000.
- [45] T. Van Cutsem, J. Howard, and M. Ribbens-Pavella, "A Two-Level Static State Estimator for Electric Power Systems," *IEEE Trans. Power App. Syst.*, no. 8, pp. 3722–3732, 1981.
- [46] L. Zhao and A. Abur, "Multi-area State Estimation using Synchronized Phasor Measurements," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 611–617, 2005.
- [47] W. Jiang, V. Vittal, and G. Heydt, "A Distributed State Estimator Utilizing Synchronized Phasor Measurements," *IEEE Trans. Power Syst.*, vol. 22, no. 2, pp. 563–571, 2007.
- [48] L. Xie, D. Choi, S. Kar, and H. Poor, "Fully Distributed State Estimation for Wide-Area Monitoring Systems," *Smart Grid, IEEE Transactions on*, vol. 3, no. 3, pp. 1154–1169, 2012.
- [49] V. Kekatos and G. Giannakis, "Distributed Robust Power System State Estimation," *Arxiv preprint arXiv:1204.0991*, 2012.
- [50] V. Blondel, J. Hendrickx, A. Olshevsky, and J. Tsitsiklis, "Convergence in Multiagent Coordination, Consensus, and Flocking," in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*. IEEE, 2005, pp. 2996–3000.
- [51] S. Strogatz, *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry, and Engineering*. Westview Pr, 1994.
- [52] R. Horn and C. Johnson, "Topics in Matrix Analysis, 1991."
- [53] K. Eriksson, D. Estep, and C. Johnson, *Applied Mathematics, Body and Soul: Derivates and Geometry in \mathbb{R}^3* . Springer Verlag, 2004, vol. 3.
- [54] A. Monticelli, "State Estimation in Electric Power Systems: A Generalized Approach, 1999."
- [55] "U. K. National Grid-Real Time Operational Data," 2009. [Online]. Available: <http://www.nationalgrid.com/uk/Electricity/Data/>
- [56] S. Salzo and S. Villa, "Convergence Analysis of a Proximal Gauss-Newton Method," *Arxiv preprint arXiv:1103.0414*, 2011.